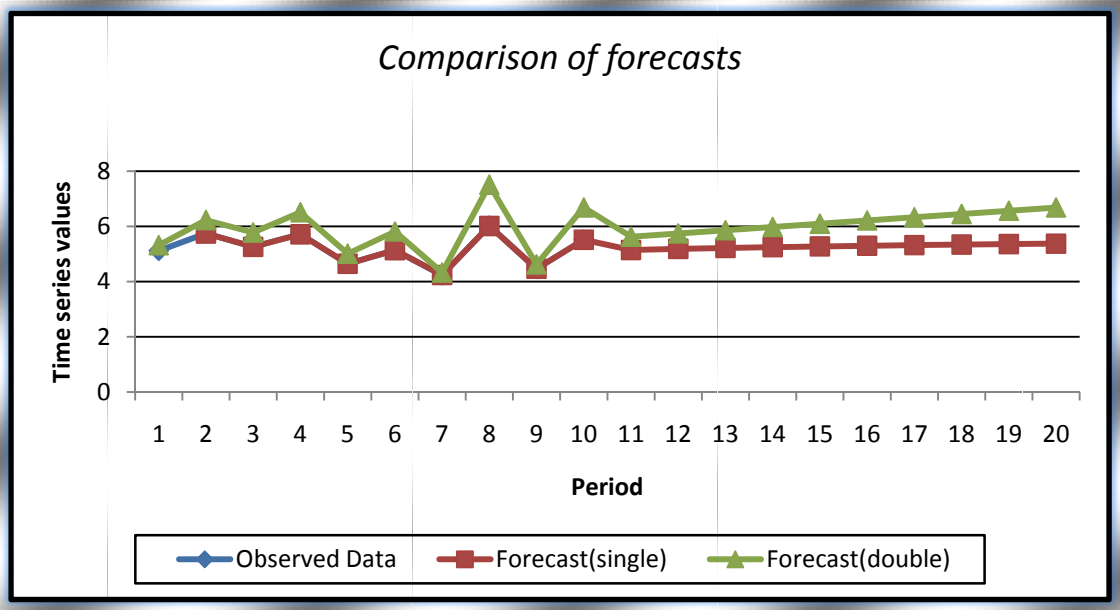
Rajesh Singh & Florentin Smarandache

editors & authors

STUDIES IN SAMPLING TECHNIQUES AND TIME SERIES ANALYSIS



Zip Publishing

2011

STUDIES IN SAMPLING TECHNIQUES AND TIME SERIES ANALYSIS

Rajesh Singh
Department of Statistics, BHU, Varanasi (U. P.), India

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

editors and authors

2011

This book can be ordered on paper or electronic formats from:

Zip Publishing
1313 Chesapeake Avenue
Columbus, Ohio 43212
USA
Tel. (614) 485-0721
E-mail: info@zippublishing.com
Website: www.zippublishing.com

Front and back covers by Editors

Peer-reviewers:

Prof. Ion Pătraşcu, Fraţii Buzeşti Natinal College, Craiova, Romania.
Prof. Luige Vlădăreanu, Institude of Solid Mechanics, Romanian Academy, Bycharest, Romania.
Eng. Victor Christianto, Malang, Indonesia.
Prof. H. P Singh, School of Studies in Statistics, Vikram University, Ujjain, M.P., India.
Dr. Jayant Singh, Department of Statistics, Rajasthan University, Jaipur, India.

ISBN: 9781599731599
Printed in the United States of America

Contents

Preface

This book has been designed for students and researchers who are working in the field of time series analysis and estimation in finite population. There are papers by Rajesh Singh, Florentin Smarandache, Shweta Maurya, Ashish K. Singh, Manoj Kr. Chaudhary, V. K. Singh, Mukesh Kumar and Sachin Malik. First chapter deals with the problem of time series analysis and the rest of four chapters deal with the problems of estimation in finite population.

The book is divided in five chapters as follows:

Chapter 1. Water pollution is a major global problem. In this chapter, time series analysis is carried out to study the effect of certain pollutants on water of Ramgarh Lake of Rajasthan, India.

Chapter 2. In this chapter family of factor-type estimators for estimating population mean of stratified population in the presence of non-response has been discussed. Choice of appropriate estimator in the family in order to get a desired level of accuracy in presence of no-response is worked out.

Chapter 3. In this chapter our aim is to discuss the existing allocation schemes in presence of non-response and to suggest some new allocation schemes utilizing the knowledge of response and non-response rates of different strata.

Chapter 4. In this chapter, we have suggested an improved estimator for estimating the population mean in stratified sampling in presence of auxiliary information.

Chapter 5. In this chapter we have proposed some estimators for the population variance of the variable under study, which make use of information regarding the population proportion possessing certain attribute.

The Editors

Time Series Analysis of Water Quality of Ramgarh Lake of Rajasthan

Rajesh Singh, Shweta Maurya
Department of Statistics, Banaras Hindu University
Varanasi-221005, INDIA

Ashish K. Singh
Raj Kumar Goel Institute of Technology, Ghaziabad, India.

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

Abstract

In this chapter an attempt has been made to study the effect of certain pollutants on water of Ramgarh Lake of Rajasthan. Time series analysis of the observed data has been done using trend, single exponential smoothing and double exponential smoothing methods.

Keywords: Pollutants, trend, single, double exponential smoothing, time series.

1. Introduction

Seventy percent of the earth's surface is covered by water. Water is undoubtedly the most precious natural resource that exists on our planet. Water is an important component of the eco-system and any imbalance created either in terms of amount which it is represent or impurities added to it, Can harm the whole eco-system. Water pollution occurs when a body of water is adversely affected due to the addition of large amount of pollutant materials to the water. When it is unfit for its intended use, water is considered polluted. There are various sources of water pollution (for detail refer to Jain (2011))
Some of the important water quality factors are:

1) Dissolved oxygen (D.O.)
2) Biological oxygen demand (B.O.D.)
3) Nitrate

4) Coliform
5) P.H.

Chemical analysis of any sample of water gives us a complete picture of its physical and chemical constituents. This will give us only certain numerical value but for estimating exact quality of water a time series system has been developed known as water quality trend, which gives us the idea of whole system for a long time (see Jain(2011)).

In this chapter we are calculating the trend values for five water parameters of Ramgarh lake in Rajasthan for the year 1995-2006 and for three parameters of Mahi river for the year 1997-2008.these methods viz. trend analysis, single smoothing are used to analyze the data.

2. Methodology

After ensuring the presence of trend in the data, smoothing of the data is the next requirement for time series analysis. For smoothing the common techniques discussed by Gardner(1985) are trend, simple exponential smoothing (SES), double exponential smoothing (DES), triple exponential smoothing (TES) and adaptive response rate simple exponential smoothing (ARRSES). Jain (2011) used trend method to analyze the data. We have extended the work of Jain (2011) and analyzed the data using SES and DES and compared these with the help of the available information. The methods are described below:

1. Fitting Of Straight line

The equation of the straight line is-

$U_t=a+b_t$

where,

u_t=observed value of the data, a=intercept value, b=slope of the straight line and

t=time (in years)

Calculation for a and b:

The normal equations for calculating a and b are

$$\sum U_t = n\,a + b\sum t$$

$$\sum t\,U_t = at + b\sum t^{**2}$$

2. Single Exponential Smoothing

The basic equation of exponential smoothing is

$$S_t = \alpha\, y_{t-1} + (1-\alpha)\, S_{t-1}, \quad 0 \le \alpha \le 1$$

and parameter α is called the smoothing constant.

Here, S_i stands for smoothed observation or EWNA and y stands for the original observation. The subscripts refer to the time periods 1,2,3…..,n.

The smoothed series starts with the smoothed version of the second observation.

3. Double Exponential Smoothing

Single smoothing does not excel in following the data when there is a trend. This situation can be improved by the introduction of a second equation with a second constant γ, which must be chosen in conjunction with α.

$$S_t = \alpha\, y_t + (1-\alpha)\,(S_{t-1} + b_{t-1}), \quad 0 \le \alpha \le 1$$

$$b_t = \gamma\,(S_t - S_{t-1}) + (1-\gamma)\, b_{t-1}, \quad 0 \le \gamma \le 1.$$

For forecasting using single and double exponential smoothing following method is used-

Forecasting with single exponential smoothing

$$S_t = \alpha y_{t-1} + (1-\alpha) S_{t-1} \quad 0 \le \alpha \le 1$$

The new forecast is the old one plus an adjustment for the error that occurred in the last forecast.

Boot strapping of forecasts

$$S_{t+1}=\alpha\, y_{origin}+(1-\alpha)\, S_t$$

This formula works when last data points and no actual observation are available**.**

Forecasting with double exponential smoothing

The one period-ahead forecast is given by:

$$F_{t+1}=S_t+b_t$$

The m-periods-ahead forecast is given by:

$$F_{t+m}=S_t+mb_t$$

(for detail of these methods refer to Gardner (1985)).

4. Results and Discussion

Table 1 shows the data on Dissolved Oxygen (D.O.) for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing.

Table 1: Data on Dissolved Oxygen (D.O.) for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing

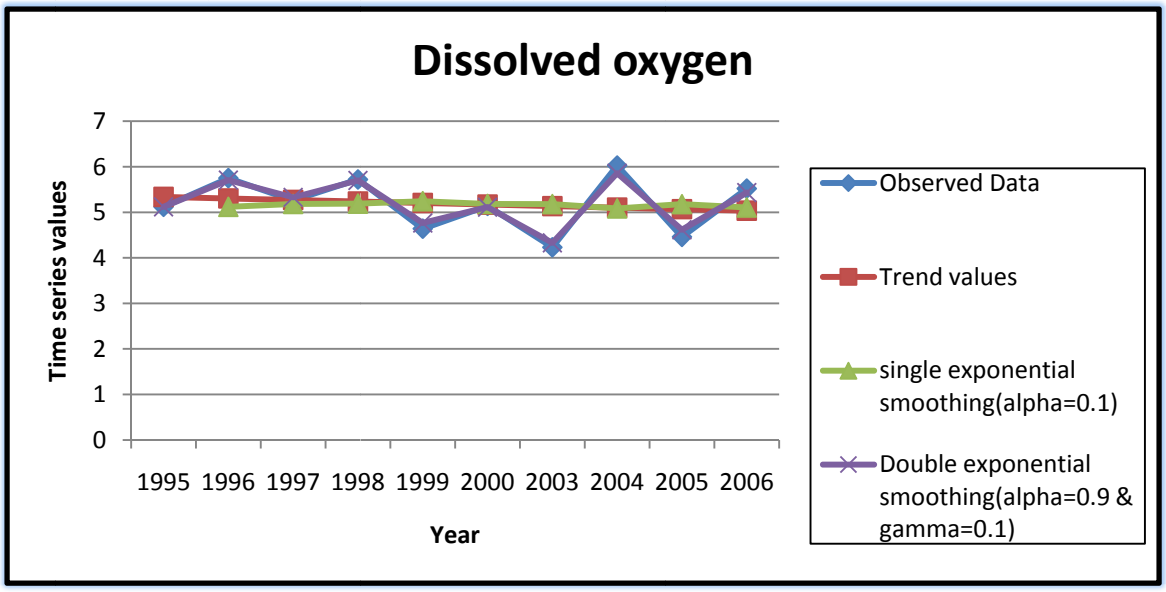
	Observed Data	**Trend values**	**Single exponential smoothing(alpha=0.1)**	**Double exponential smoothing(alpha=0.9 and gamma=0.1)**
1995	5.12	5.3372		5.12
1996	5.75	5.3036	5.12	5.707
1997	5.26	5.27	5.183	5.32857
1998	5.72	5.2364	5.1907	5.698556
1999	4.64	5.2028	5.24363	4.765484
2000	5.14	5.1692	5.183267	5.110884
2003	4.23	5.1356	5.17894	4.329044
2004	6.026	5.102	5.084046	5.858346
2005	4.46	5.0684	5.178242	4.616965
2006	5.52	5.0348	5.106417	5.4327
Total	51.866	51.86	46.46824	51.96755

For trend, the fitted line is

$U_t = 5.1866-0.0169 * t$

with mean squared error (MSE) 0.3077. For single exponential smoothing various values of α are tried and minimum MSE = 0.3915 was obtained for α = 0.1. For smoothing of the data, Holt's double exponential smoothing was found to be most appropriate. Various combinations of α and γ both ranging between 0.1 and 0.9 with increments of 0.1 were tried and MSE = 0.0094 was least for α = 0.9 and γ = 0.1.

Figure 1- Graph of observed data and fitted values using trend, single and double exponential smoothing of Dissolved Oxygen (D.O.) for Ramgardh Lake for the years 1995-2006



Adequate dissolved oxygen is necessary for good water quality. Oxygen is a necessary element to all forms of life. Natural stream purification processes require adequate oxygen levels in order to provide for aerobic life forms. As dissolved oxygen levels in water drop below 5.0 mg/l, aquatic life is put under stress (for details see www.state.ky.us).

Form Table 1 and Figure 1, we observe that level of D.O. in Ramgarh Lake was above the required standard 5.0 mg/l, except for the two years 1999 and 2005.

Table 2: Comparison of forecasts

Period	Observed Data	Forecast(single)	Forecast(double)
1	5.12		5.32
2	5.75	5.7437	6.23084
3	5.26	5.25923	5.77869651
4	5.72	5.714707	6.510832048
5	4.64	4.6460363	5.013406419
6	5.14	5.14043267	5.815207177
7	4.23	4.239489403	4.32655631
8	6.026	6.016580463	7.511558059
9	4.46	4.467182416	4.621632535
10	5.52	5.515864175	6.686376834
11		5.147775731	5.6266
12		5.184998158	5.7442
13		5.218498342	5.8618
14		5.248648508	5.9794
15		5.275783657	6.097
16		5.300205291	6.2146
17		5.322184762	6.3322
18		5.341966286	6.4498
19		5.359769657	6.5674
20		5.375792692	6.685

Figure 2: Graph of forecasts

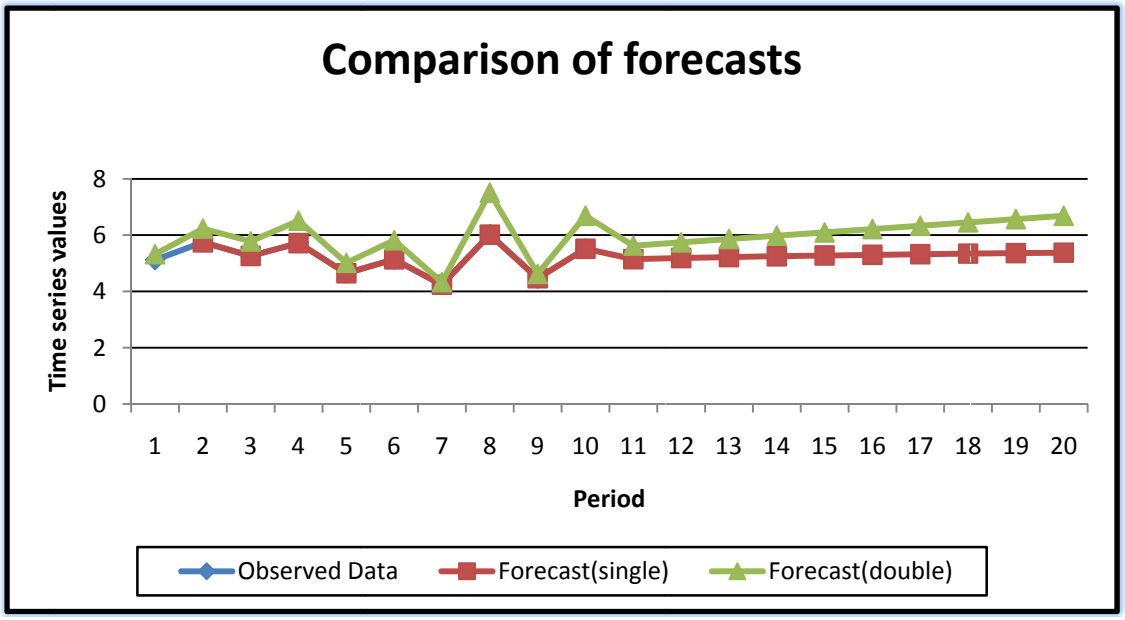


Table 3 shows the data on Nitrate for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing.

Table 3: Data on Nitrate for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing

Year(x)	Observed Data	Trend values	Single exponential smoothing(alpha=0.1)	Double exponential smoothing(alpha=0.9 and gamma=0.1)
1995	0.32	0.588		0.32
1996	1.28	0.546	0.32	1.176
1997	0.38	0.504	0.416	0.46096
1998	0.08	0.462	0.4124	0.104883
1999	0.04	0.42	0.37916	0.028797
2000	0.92	0.378	0.345244	0.815205
2003	0.24	0.336	0.40272	0.300708
2004	0.25	0.294	0.386448	0.247331
2005	0.186	0.252	0.372803	0.184874
2006	0.3	0.21	0.354123	0.281431
Total	3.996	3.99	3.388897	3.920189

For trend, the fitted line is

$U_t = 0.3996 - 0.0216 * t$

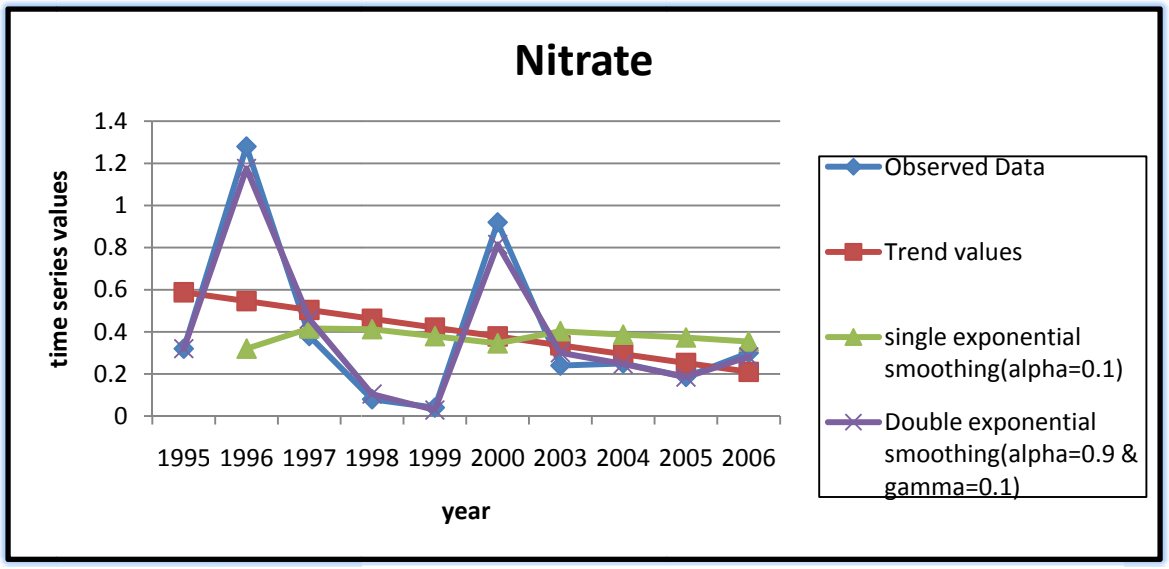
with MSE= 0.3078. For single exponential smoothing various values of α are tried and minimum MSE = 0.3412 was obtained for $\alpha = 0.1$. For smoothing of the data, Holt's double exponential smoothing was found to be most appropriate. Various combinations of α and γ both ranging between 0.1 and 0.9 with increments of 0.1 were tried and MSE = 0.0033 was least for $\alpha = 0.9$ and $\gamma = 0.1$.

Nitrites can produce a serious condition in fish called "brown blood disease." Nitrites also react directly with hemoglobin in human blood and other warm-blooded animals to produce methemoglobin. Methemoglobin destroys the ability of red blood cells to transport oxygen. This condition is especially serious in babies under three months of age. It causes a condition known as methemoglobinemia or "blue baby" disease. Water with nitrite levels exceeding 1.0 mg/l should not be used for feeding

babies. Nitrite/nitrogen levels below 90 mg/l and nitrate levels below 0.5 mg/l seem to have no effect on warm water fish (for details see www.state.ky.us).

Form Table 3 and Figure 3, we observe that level of Nitrate in Ramgarh Lake was below the standard 1.0 mg/l, except for the year 1996.

Figure 3: Graph of observed data and fitted values using trend, single and double exponential smoothing of Nitrate for Ramgardh lake for the years 1995-2006



Nitrites can produce a serious condition in fish called "brown blood disease." Nitrites also react directly with hemoglobin in human blood and other warm-blooded animals to produce methemoglobin. Methemoglobin destroys the ability of red blood cells to transport oxygen. This condition is especially serious in babies under three months of age. It causes a condition known as methemoglobinemia or "blue baby" disease. Water with nitrite levels exceeding 1.0 mg/l should not be used for feeding babies. Nitrite/nitrogen levels below 90 mg/l and nitrate levels below 0.5 mg/l seem to have no effect on warm water fish (for details see www.state.ky.us).

Form Table 3 and Figure 3, we observe that level of Nitrate in Ramgarh Lake was below the standard 1.0 mg/l, except for the year 1996.

Table 4: Comparison of forecasts

Period	Observed Data	Forecast(single)	Forecast(double)
1	0.32	0.416	0.24
2	1.28	0.4124	1.1896
3	0.38	0.37916	0.328832
4	0.08	0.345244	-0.07203
5	0.04	0.40272	-0.12795
6	0.92	0.386448	0.847085
7	0.24	0.372803	0.223314
8	0.25	0.354123	0.17474
9	0.186	0.34871	0.114309
10	0.3	0.348711	0.244291
11		0.34384	0.24426
12		0.339456	0.20712
13		0.33551	0.16998
14		0.331959	0.13284
15		0.328763	0.0957
16		0.325887	0.05856
17		0.323298	0.02142
18		0.320968	-0.01572
19		0.318872	-0.05286
20		0.316984	-0.09

Figure 4 : Graph of forecasts

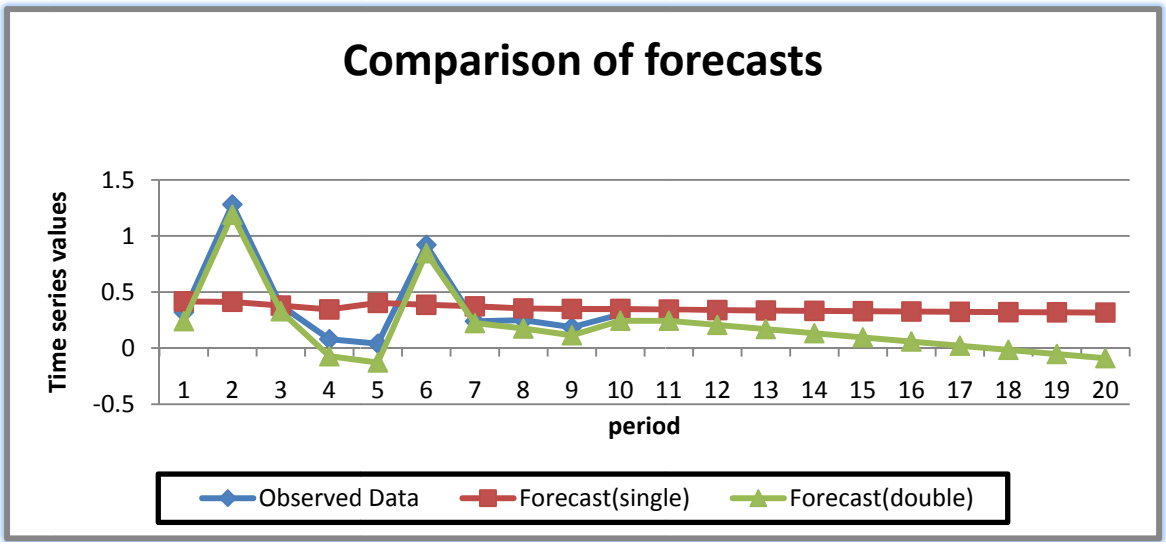


Table 5 shows the data on B.O.D. for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing.

Table 5: Data on Biological oxygen demand (B.O.D.) for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing

Year(x)	Observed Data	Trend values	single exponential smoothing(alpha=0.1)	Double exponential smoothing(alpha=0.9 & gamma=0.1)
1995	1.96	5.122		1.96
1996	14.09	4.762	1.96	12.87167
1997	1.84	4.402	3.173	3.047483
1998	1.8	4.042	3.0397	1.920392
1999	1.46	3.682	2.91573	1.490847
2000	2.78	3.322	2.770157	2.633116
2003	2.76	2.962	2.771141	2.742563
2004	1.976	2.602	2.770027	2.049477
2005	2.78	2.242	2.690624	2.697155
2006	3.58	1.882	2.699562	3.489379
Total	35.026	35.02	24.78994	34.90208

Biochemical oxygen demand is a measure of the quantity of oxygen used by microorganisms (e.g., aerobic bacteria) in the oxidation of organic matter. Natural sources of organic matter include plant decay and leaf fall. However, plant growth and decay may be unnaturally accelerated when nutrients and sunlight are overly abundant due to human influence. Urban runoff carries pet wastes from streets and sidewalks; nutrients from lawn fertilizers; leaves, grass clippings, and chapter from residential areas, which increase oxygen demand. Oxygen consumed in the decomposition process robs other aquatic organisms of the oxygen they need to live. Organisms that are more tolerant of lower dissolved oxygen levels may replace a diversity of natural water systems contain bacteria, which need oxygen (aerobic) to survive. Most of them feed on dead algae and other dead organisms and are part of the decomposition cycle. Algae and other producers in the water take up inorganic nutrients and use them in the process of building up their organic tissues (for details refer to www.freedrinkingwater.com).

For trend, the fitted line is

$$U_t = 3.5026+0.18094 * t$$

with MSE= 11.74366. For single exponential smoothing various values of α are tried and minimum MSE = 17.1093 was obtained for α = 0.1. For smoothing of the data, Holt's double exponential smoothing was found to be most appropriate. Various combinations of α and γ both ranging between 0.1 and 0.9 with increments of 0.1 were tried and MSE = 0.3000 was least for α = 0.9 and γ = 0.1.

Figure 5: Graph of observed data and fitted values using trend, single and double exponential smoothing of B.O.D. for Ramgardh Lake for the years 1995-2006

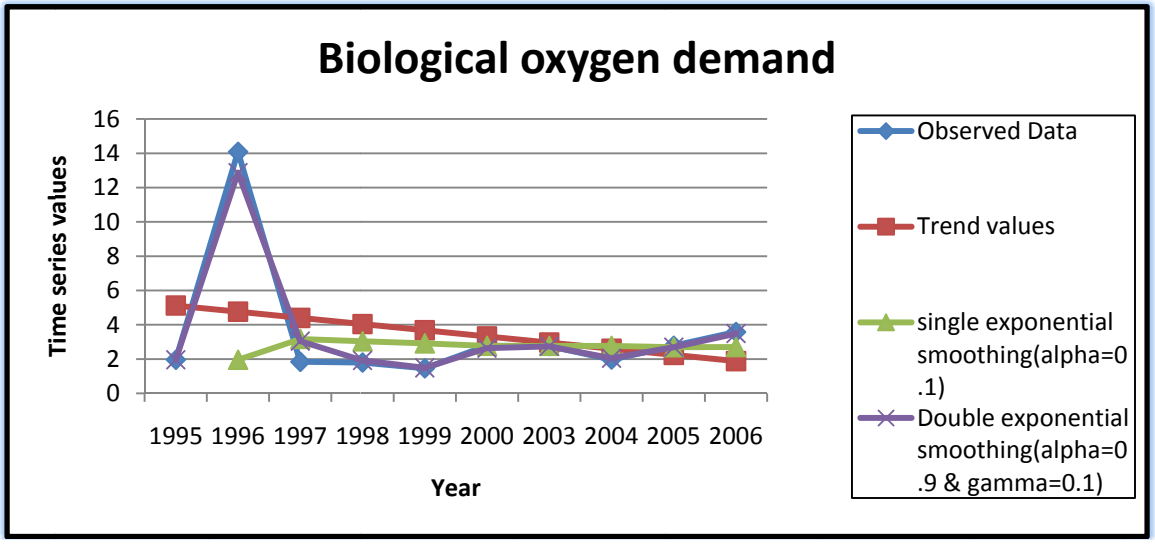


Table 6: Comparison of forecasts

Period	Observed Data	Forecast(single)	Forecast(double)
1	1.96		1.906667
2	14.09	3.173	13.91483
3	1.84	3.0397	3.003915
4	1.8	2.91573	1.768471
5	1.46	2.770157	1.311164
6	2.78	2.771141	2.585629
7	2.76	2.770027	2.710768
8	1.976	2.690624	1.951553
9	2.78	2.699562	2.673792
10	3.58	2.787606	3.547575
11		2.7871	3.5474
12		2.86639	3.6055
13		2.937751	3.6636
14		3.001976	3.7217
15		3.059778	3.7798
16		3.1118	3.8379
17		3.15862	3.896
18		3.200758	3.9541
19		3.238683	4.0122
20		3.272814	4.0703

Figure 6 : Graph of forecasts

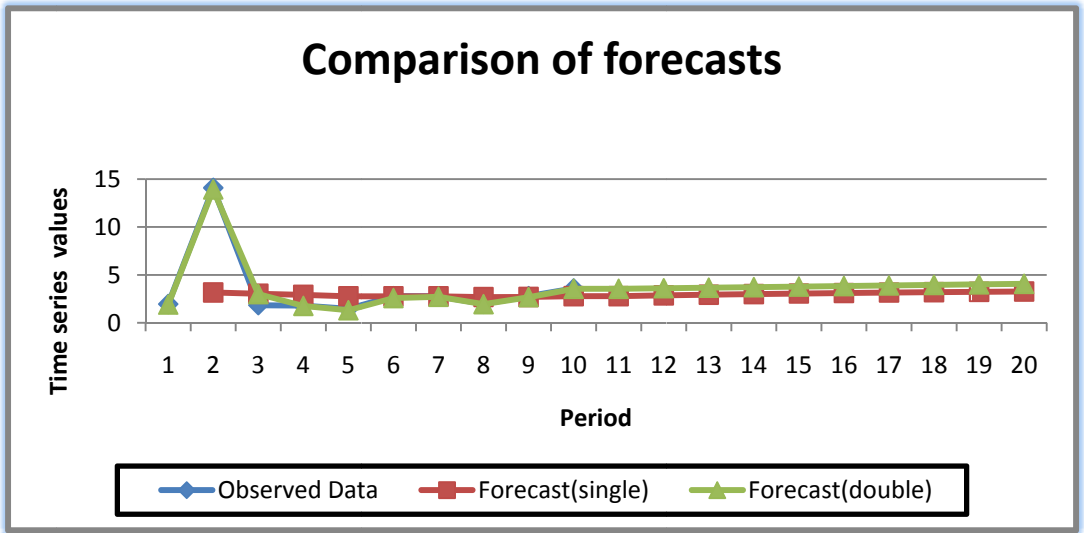


Table 7 shows the data on Total Colifirm for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing.

Table 7 : Data on Total Colifirm for Ramgardh lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing

Year(x)	Observed Data	Trend values	single exponential smoothing(alpha=0.6)	Double exponential smoothing(alpha=0.9 & gamma=0.1)
1995	1169	687.894		1169
1996	285	615.314	1169	339.2333
1997	840.75	542.734	638.6	751.5507
1998	144	470.154	759.89	173.7353
1999	65.33	397.574	390.356	42.47463
2000	121.5	324.994	195.3404	81.95854
2003	119	252.414	151.0362	87.21566
2004	720.6	179.834	131.8145	632.042
2005	86	107.254	485.0858	123.3548
2006	61.66	34.674	245.6343	47.21817
Total	3612.84	3612.84	4166.757	3447.783

Total coliform bacteria are a collection of relatively harmless microorganisms that live in large numbers in the intestines of man and warm- and cold-blooded animals. They aid in the digestion of food. A specific subgroup of this collection is the fecal coliform bacteria, the most common member being Escherichia coli. These organisms may be separated from the total coliform group by their ability to grow at elevated temperatures and are associated only with the fecal material of warm-blooded animals. The presence of fecal coliform bacteria in aquatic environments indicates that the water has been contaminated with the fecal material of man or other animals. The presence of fecal contamination is an indicator that a potential health risk exists for individuals exposed to this water (for details see www.state.ky.us).

For trend, the fitted line is

$U_t = 361.284 - 36.2989^* t$

with MSE= 99896.33. For single exponential smoothing various values of α are tried and minimum MSE = 205949.6 was obtained for α = 0.6. For smoothing of the data, Holt's double exponential smoothing was found to be most appropriate. Various combinations of α and γ both ranging between 0.1 and 0.9 with increments of 0.1 were tried and MSE = 2432.458 was least for α = 0.9 and γ = 0.1.

Figure 7: Graph of observed data and fitted values using trend, single and double exponential smoothing of Total Coliform for Ramgardh Lake for the years 1995-2006.

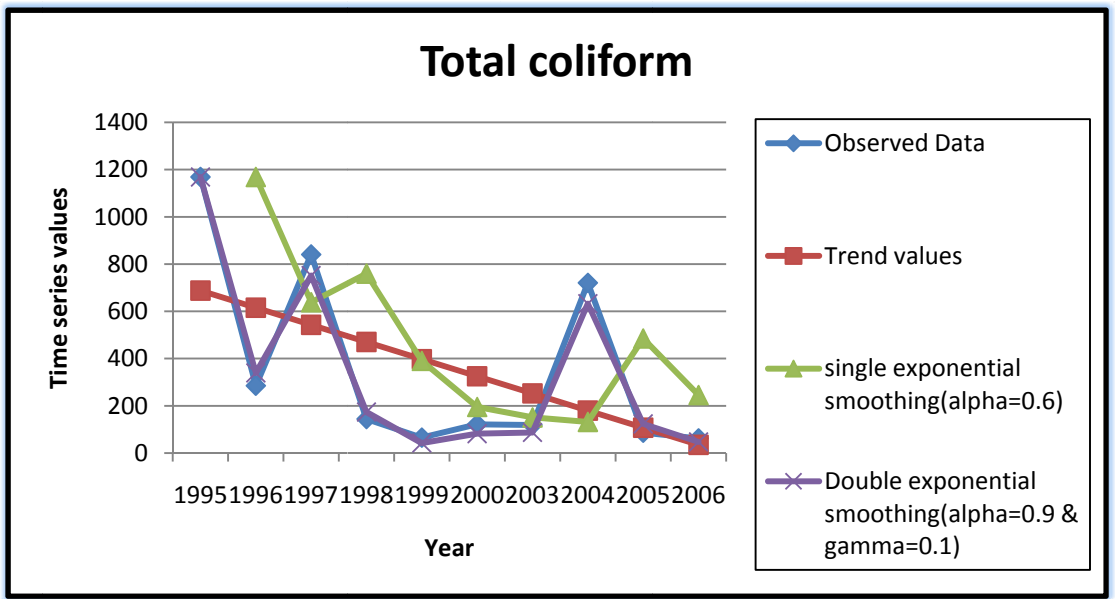


Table 8: Comparison of forecasts

Period	Observed Data	Forecast(single)	Forecast(double)
1	1169		827.3333
2	285	638.6	-51.2433
3	840.75	759.89	441.3534
4	144	390.356	-163.224
5	65.33	195.3404	-273.915
6	121.5	151.0362	-198.843
7	119	131.8145	-164.98
8	720.6	485.0858	459.5482
9	86	245.6343	-82.7583
10	61.66	135.2497	-145.897
11		135.248	-145.897
12		91.0952	-339.012
13		73.43408	-532.127
14		66.36963	-725.242
15		63.54385	-918.357
16		62.41354	-1111.47
17		61.96142	-1304.59
18		61.78057	-1497.7
19		61.70823	-1690.82
20		61.67929	-1883.93

Figure 8: Comparison of forecasts

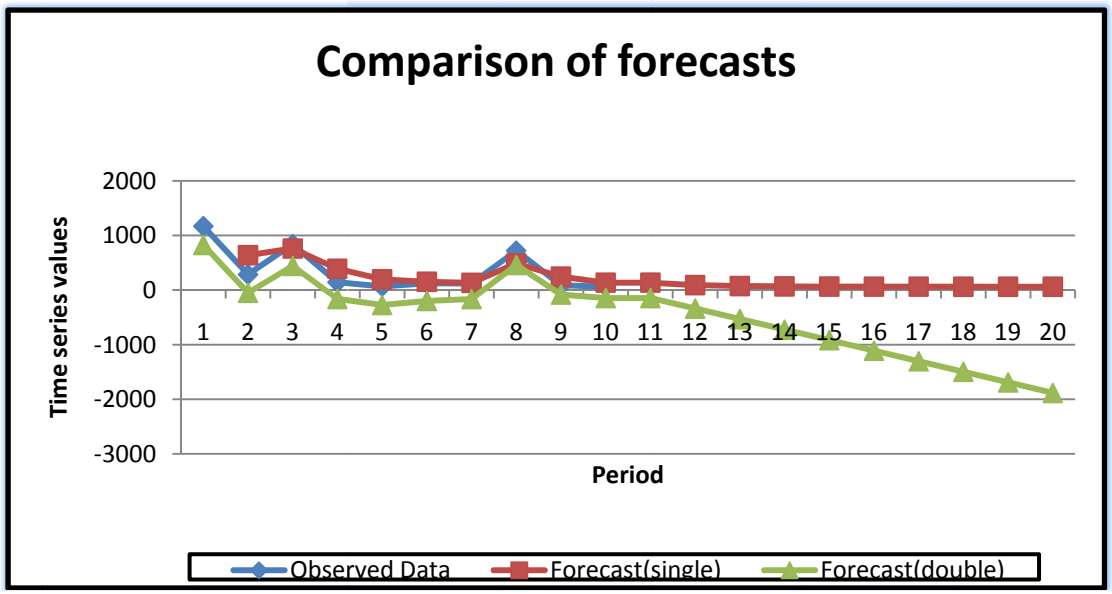


Table 9 shows the data on pH for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing.

Table 9: Data on pH for Ramgardh Lake for the years 1995-2006 and fitted values using trend, single and double exponential smoothing

Year(x)	Observed Data	Trend values	single exponential smoothing(alpha=0.2)	Double exponential smoothing(alpha=0.9 & gamma=0.1)
1995	7.64	6.29		7.64
1996	7.48	6.65	7.64	7.509667
1997	7.87	7.01	7.608	7.844963
1998	8.05	7.37	7.6604	8.042746
1999	8.44	7.73	7.73832	8.414177
2000	8.3	8.09	7.878656	8.327645
2003	7.25	8.45	7.962925	7.371503
2004	8.28	8.81	7.82034	8.191954
2005	7.87	9.17	7.912272	7.912923
2006	8.006	9.53	7.903817	8.003557
Total	79.186	79.1	70.12473	79.25914

pH is a measure of the acidic or basic (alkaline) nature of a solution. The concentration of the hydrogen ion [H+] activity in a solution determines the pH. A pH range of 6.0 to 9.0 appears to provide protection for the life of freshwater fish and bottom

dwelling invertebrates. The most significant environmental impact of pH involves synergistic effects. Synergy involves the combination of two or more substances which produce effects greater than their sum. This process is important in surface waters. Runoff from agricultural, domestic, and industrial areas may contain iron, aluminum, ammonia, mercury or other elements. The pH of the water will determine the toxic effects, if any, of these substances. For example, 4 mg/l of iron would not present a toxic effect at a pH of 4.8. However, as little as 0.9 mg/l of iron at a pH of 5.5 can cause fish to die (for details see www.state.ky.us).

For trend, the fitted line is

$U_t = 7.9186 + 0.1845 * t$

with MSE = 0.9995. For single exponential smoothing various values of α are tried and minimum MSE = 0.1831 was obtained for α = 0.2. For smoothing of the data, Holt's double exponential smoothing was found to be most appropriate. Various combinations of α and γ both ranging between 0.1 and 0.9 with increments of 0.1 were tried and MSE = 0.002735was least for α = 0.9 and γ = 0.1.

Figure 9: Graph of observed data and fitted values using trend, single and double exponential smoothing of pH for Ramgardh Lake for the years 1995-2006.

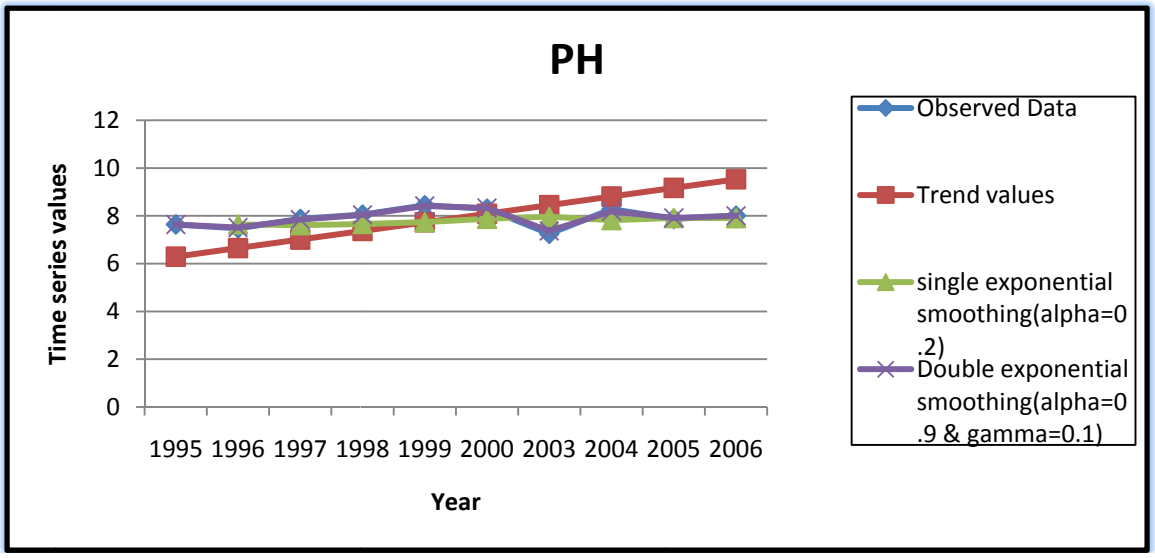
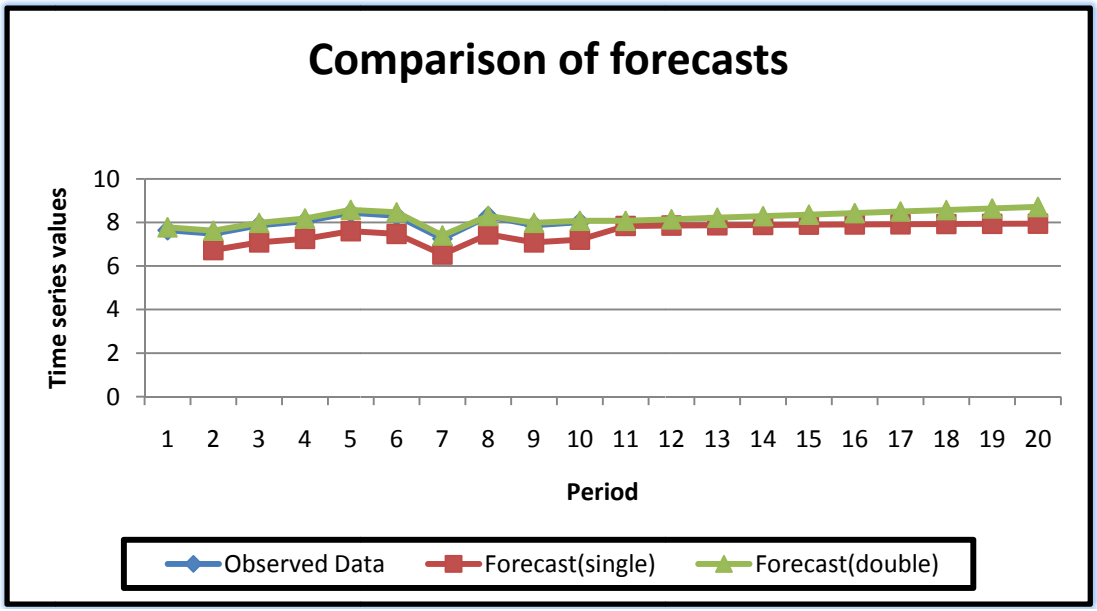


Table 10: Comparison of forecasts

Period	Observed Data	Forecast(single)	Forecast(double)
1	7.64		7.776667
2	7.48	6.732	7.619633
3	7.87	7.083	7.977463
4	8.05	7.245	8.181774
5	8.44	7.596	8.576446
6	8.3	7.47	8.465033
7	7.25	6.525	7.399539
8	8.28	7.452	8.299231
9	7.87	7.083	7.981569
10	8.006	7.2054	8.074402
11		7.8368	8.074395
12		7.85372	8.14524
13		7.868948	8.216085
14		7.882653	8.28693
15		7.894988	8.357775
16		7.906089	8.42862
17		7.91608	8.499465
18		7.925072	8.57031
19		7.933165	8.641155
20		7.940448	8.712

Figure 10: Graph of forecasts



We observe from the calculations for the different parameters of pollutants that double smoothing follows the data much closer than single smoothing. Furthermore, for forecasting single smoothing cannot do better than projecting a straight horizontal line, which is not very likely to occur in reality. So for forecasting purposes for our data double exponential smoothing is more preferable.

Conclusion

From the above discussions, we observe that the various pollutants considered in the article may have very hazardous effect on quality of water. Increase of pollutants in water beyond a certain limit may be dangerous for aquatic animals. Also, according to recent reports, most of the tap and well water in the India are not safe for drinking due to presence of various pollutants in inappropriate percentage. Now, we have reached the point where all sources of our drinking water, including municipal water systems, wells, lakes, rivers, and even glaciers, contain some level of contamination. So, we need to keep a routine check of the quality of water so that we can lead a healthy life.

References

Gardner, E. S. (1985) : Exponential smoothing- The state of the art. Jour. Of Forecasting,4, 1-28.

Jain, Smita (2011): A Statistical study of effect of different pollutants on water (with specific reference to Rajasthan). Unpublished thesis submitted to University of Rajasthan, India.

www.state.ky.us

www.freedrinkingwater.com

Estimating the Population Mean in Stratified Population using Auxiliary Information under Non-Response

Manoj Kr. Chaudhary, V. K. Singh and Rajesh Singh
Department of Statistics, Banaras Hindu University
Varanasi-221005, INDIA

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

Abstract

The present chapter deals with the study of general family of factor-type estimators for estimating population mean of stratified population in the presence of non-response whenever information on an auxiliary variable are available. The proposed family includes separate ratio, product, dual to ratio and usual sample mean estimators as its particular cases and exhibits some nice properties as regards to locate the optimum estimator belonging to the family. Choice of appropriate estimator in the family in order to get a desired level of accuracy even if non-response is high, is also discussed. The empirical study has been carried out in support of the results.

Keywords: Factor-type estimators, Stratified population, Non-response, Optimum estimator, Empirical study.

1. Introduction

In sampling theory the use of suitable auxiliary information results in considerable reduction in variance of the estimator. For this reason, many authors used the auxiliary information at the estimation stage. Cochran (1940) was the first who used the auxiliary information at the estimation stage in estimating the population parameters. He proposed the ratio estimator to estimate the population mean or total of a character under study. Hansen *et. al.* (1953) suggested the difference estimator which was subsequently modified to give the linear regression estimator for the population mean or its total. Murthy (1964) have studied the product estimator to estimate the population mean or total when the character under study and the auxiliary character are negatively

correlated. These estimators can be used more efficiently than the mean per unit estimator.

There are several authors who have suggested estimators using some known population parameters of an auxiliary variable. Upadhyaya and Singh (1999) have suggested the class of estimators in simple random sampling. Kadilar and Cingi (2003) and Shabbir and Gupta (2005) extended these estimators for the stratified random sampling. Singh et. al. (2008) suggested class of estimators using power transformation based on the estimators developed by Kadilar and Cingi (2003). Kadilar and Cingi (2005) and Shabbir and Gupta (2006) have suggested new ratio estimators in stratified sampling to improve the efficiency of the estimators. Koyuncu and Kadilar (2008) have proposed families of estimators for estimating population mean in stratified random sampling by considering the estimators proposed in Searls (1964) and Khoshnevisan et. al. (2007). Singh and Vishwakarma (2008) have suggested a family of estimators using transformation in the stratified random sampling. Recently, Koyuncu and Kadilar (2009) have proposed a general family of estimators, which uses the information of two auxiliary variables in the stratified random sampling to estimate the population mean of the variable under study.

The works which have been mentioned above are based on the assumption that both the study and auxiliary variables are free from any kind of non-sampling error. But, in practice, however the problem of non-response often arises in sample surveys. In such situations while single survey variable is under investigation, the problem of estimating population mean using sub-sampling scheme was first considered by Hansen and Hurwitz (1946). If we have incomplete information on study variable X_0 and complete information on auxiliary variable X_1, in other words if the study variable is affected by non-response error but the auxiliary variable is free from non-response. Then utilizing the Hansen-Hurwitz (1946) technique of sub-sampling of the non-respondents, the conventional ratio and product estimators in the presence of non-response are respectively given by

$$T_{0R}^{*} = \left(T_{0HH} / \bar{x}_1\right)\overline{X}_1 \tag{1.1}$$

and $$T_{0P}^{*} = \left(T_{0HH}.\bar{x}_1\right)/\overline{X}_1 . \qquad (1.2)$$

The purpose of the present chapter is to suggest separate-type estimators in stratified population for estimating population mean using the concept of sub-sampling of non-respondents in the presence of non-response in study variable in the population. In this context, the information on an auxiliary characteristic closely related to the study variable, has been utilized assuming that it is free from non-response.

In order to suggest separate-type estimators, we have made use of Factor-Type Estimators (FTE) proposed by Singh and Shukla (1987). FTE define a class of estimators involving usual sample mean estimator, usual ratio and product estimators and some other estimators existing in literature. This class of estimators exhibits some nice properties which have been discussed in subsequent sections.

2. Sampling Strategy and Estimation Procedure

Let us consider a population consisting of N units divided into k strata. Let the size of i^{th} stratum is N_i, $(i = 1,2,...,k)$ and we decide to select a sample of size n from the entire population in such a way that n_i units are selected from the i^{th} stratum. Thus, we have $\sum_{i=1}^{k} n_i = n$. Let the non-response occurs in each stratum. Then using Hansen and Hurwitz (1946) procedure we select a sample of size m_i units out of n_{i2} non-respondent units in the i^{th} stratum with the help of simple random sampling without replacement (SRSWOR) scheme such that $n_{i2} = L_i m_i$, $L_i \geq 1$ and the information are observed on all the m_i units by interview method.

The Hansen-Hurwitz estimator of population mean $\overline{X}_{0i}$ of study variable X_0 for the i^{th} stratum will be

$$T_{0i}^{*} = \frac{n_{i1}\bar{x}_{0i1} + n_{i2}\bar{x}_{0mi}}{n_i}, \qquad \left(i = 1,2,...,k\right) \qquad (2.1)$$

where $\bar{x}_{0i1}$ and $\bar{x}_{0mi}$ are the sample means based on n_{i1} respondent units and m_i non-respondent units respectively in the i^{th} stratum for the study variable.

Obviously T_{0i}^{*} is an unbiased estimator of $\overline{X}_{0i}$. Combining the estimators over all the strata we get the estimator of population mean $\overline{X}_0$ of study variable X_0, given by

$$T_{0st}^{*} = \sum_{i=1}^{k} p_i T_{0i}^{*} \tag{2.2}$$

where $p_i = \frac{N_i}{N}$.

which is an unbiased estimator of $\overline{X}_0$. Now, we define the estimator of population mean $\overline{X}_1$ of auxiliary variable X_1 as

$$T_{1st} = \sum_{i=1}^{k} p_i \bar{x}_{1i} \tag{2.3}$$

where $\bar{x}_{1i}$ is the sample mean based on n_i units in the i^{th} stratum for the auxiliary variable. It can easily be seen that T_{1st} is an unbiased estimator of $\overline{X}_1$ because $\bar{x}_{1i}$ gives unbiased estimates of the population mean $\overline{X}_{1i}$ of auxiliary variable for the i^{th} stratum.

3. Suggested Family of Estimators

Let us now consider the situation in which the study variable is subjected to non-response and the auxiliary variable is free from non-response. Motivated by Singh and Shukla (1987), we define the separate-type family of estimators of population mean $\overline{X}_0$ using factor-type estimators as

$$T_{FS}(\alpha) = \sum_{i=1}^{k} p_i T_{Fi}^{*}(\alpha) \tag{3.1}$$

where $$T_{Fi}^{*}(\alpha) = T_{0i}^{*}\left[\frac{(A+C)\overline{X}_{1i} + fB\bar{x}_{1i}}{(A+fB)\overline{X}_{1i} + C\bar{x}_{1i}}\right] \tag{3.2}$$

and $f = \frac{n}{N}$, $A = (\alpha-1)(\alpha-2)$, $B = (\alpha-1)(\alpha-4)$, $C = (\alpha-2)(\alpha-3)(\alpha-4)$; $\alpha > 0$.

3.1 Particular Cases of $T_{FS}(\alpha)$

Case-1: If $\alpha = 1$ then $A = B = 0$, $C = -6$

so that $T_{Fi}^{*}(1) = T_{0i}^{*}\frac{\overline{X}_{1i}}{\bar{x}_{1i}}$

and hence $T_{FS}(1) = \sum_{i=1}^{k} p_i T_{0i}^{*}\frac{\overline{X}_{1i}}{\bar{x}_{1i}}$. (3.3)

Thus, $T_{FS}(1)$ is the usual separate ratio estimator under non-response.

Case-2: If $\alpha = 2$ then $A = 0 = C$, $B = -2$

so that $T_{Fi}^{*}(2) = T_{0i}^{*}\frac{\bar{x}_{1i}}{\overline{X}_{1i}}$

and hence $T_{FS}(2) = \sum_{i=1}^{k} p_i T_{0i}^{*}\frac{\bar{x}_{1i}}{\overline{X}_{1i}}$ (3.4)

which is the usual separate product estimator under non-response.

Case-3: If $\alpha = 3$ then $A = 2$, $B = -2$, $C = 0$

so that $T_{Fi}^{*}(3) = T_{0i}^{*}\frac{\overline{X}_{1i} - f\bar{x}_{1i}}{(1-f)\overline{X}_{1i}}$

and hence $T_{FS}(3) = \sum_{i=1}^{k} p_i T_{Fi}^{*}(3)$ (3.5)

which is the separate dual to ratio-type estimator under non-response. The dual to ratio type estimator was proposed by Srivenkataramana (1980).

Case-4: If $\alpha = 4$ then $A = 6$, $B = 0$, $C = 0$

so that $T_{Fi}^{*}(4) = T_{0i}^{*}$

and hence $T_{FS}(4) = \sum_{i=1}^{k} p_i T_{0i}^{*} = T_{0st}^{*}$ (3.6)

which is usual mean estimator defined in stratified population under non-response.

3.2 Properties of $T_{FS}(\alpha)$

Using large sample approximation, the bias of the estimator $T_{FS}(\alpha)$, up to the first order of approximation was obtained following Singh and Shukla (1987) as

$$B\left[T_{FS}(\alpha)\right] = E\left[T_{FS}(\alpha) - \overline{X}_0\right]$$

$$= \phi(\alpha)\sum_{i=1}^{k} p_i \overline{X}_{0i}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left[\frac{C}{A + fB + C}C_{1i}^{2} - \rho_{01i}C_{0i}C_{1i}\right] \quad (3.7)$$

where $\phi(\alpha) = \frac{C - fB}{A + fB + C}$, $C_{0i} = \frac{S_{0i}}{\overline{X}_{0i}}$, $C_{1i} = \frac{S_{1i}}{\overline{X}_{1i}}$, S_{0i}^{2} and S_{1i}^{2} are the population mean squares of study and auxiliary variables respectively in the i^{th} stratum . ρ_{01i} is the population correlation coefficient between X_0 and X_1 in the i^{th} stratum. The Mean Square Error (MSE) up to the first order of approximation was derived as

$$M\left[T_{FS}(\alpha)\right] = E\left[T_{FS}(\alpha) - \overline{X}_0\right]^2$$

$$= \sum_{i=1}^{k} p_i^2 MSE\left[T_{Fi}^{*}(\alpha)\right]$$

$$= \sum_{i=1}^{k} p_i^2 \overline{X}_{0i}^2 \left[\frac{V\left(T_{0i}^*\right)}{\overline{X}_{0i}^2} + \phi^2(\alpha) \ \frac{V\left(\bar{x}_{1i}\right)}{\overline{X}_{1i}^2} - 2\phi(\alpha) \frac{Cov\left(T_{0i}^*, \bar{x}_{1i}\right)}{\overline{X}_{0i}\overline{X}_{1i}} \right].$$

Since $V\left(T_{0i}^*\right) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{0i}^2 + \frac{L_i - 1}{n_i} W_{i2} S_{0i2}^2$, $V\left(\bar{x}_{1i}\right) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{1i}^2$

and $Cov\left(T_{0i}^*, \bar{x}_{1i}\right) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \rho_{01i} S_{0i} S_{1i}$ [due to Singh (1998)].

where S_{0i2}^2 is the population mean square of the non-response group in the i^{th} stratum and W_{i2} is the non-response rate of the i^{th} stratum in the population.

Therefore, we have

$$M\left[T_{FS}(\alpha)\right] = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \left[S_{0i}^2 + \varphi(\alpha)^2 R_{01i}^2 S_{1i}^2 - 2\varphi(\alpha) R_{01i} \rho_{01i} S_{0i} S_{1i}\right]$$

$$+ \sum_{i=1}^{k} \frac{L_i - 1}{n_i} W_{i2} p_i^2 S_{0i2}^2 \qquad (3.8)$$

where $R_{01i} = \frac{\overline{X}_{0i}}{\overline{X}_{1i}}$.

3.3 Optimum Choice of α

In order to obtain minimum MSE of $T_{FS}(\alpha)$, we differentiate the MSE with respect to α and equate the derivative to zero

$$\sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \left[2\phi'(\alpha)\phi(\alpha) R_{01i}^2 S_{1i}^2 - 2\phi'(\alpha) R_{01i} \rho_{01i} S_{0i} S_{1i}\right] = 0, \qquad (3.9)$$

where $\phi'(\alpha)$ stands for first derivative of $\phi(\alpha)$.From the above expression, we have

$$\phi(\alpha)=\frac{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2R_{01i}\rho_{01i}S_{0i}S_{1i}}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2R_{01i}^2S_{1i}^2}=V \text{ (say).} \qquad (3.10)$$

It is easy to observe that $\phi(\alpha)$ is a cubic equation in the parameter α. Therefore, the equation (3.10) will have at the most three real roots at which the MSE of the estimator $T_{FS}(\alpha)$ attains its minimum.

Let the equation (3.10) yields solutions as α_0, α_1 and α_2 such that $M[T_{FS}(\alpha)]$ is same. A criterion of making a choice between α_0, α_1 and α_2 is that "compute the bias of the estimator at $\alpha=\alpha_0$, α_1 and α_2 and select α_{opt} at which bias is the least". This is a novel property of the FTE.

3.4 Reducing MSE through Appropriate Choice of α

By using FTE for defining the separate-type estimators in this chapter, we have an advantage in terms of the reduction of the value of MSE of the estimator to a desired extent by an appropriate choice of the parameter α even if the non-response rate is high in the population. The procedure is described below:

Since MSE's of the proposed strategies are functions of the unknown parameter α as well as functions of non-response rates W_{i2}, it is obvious that if α is taken to be constant, MSE's increase with increasing non-response rate, if other characteristics of the population remain unchanged, along with the ratio to be sub sampled in the non-response class, that is, L_i. It is also true that more the non-response rate, greater would be the size of the non-response group in the sample and, therefore, in order to lowering down the MSE of the estimator, the size of sub sampled units should be increased so as to keep the value of L_i in the vicinity of 1; but this would, in term, cost more because more effort and money would be required to obtain information on sub sampled units through personal interview method. Thus, increasing the size of the sub sampled units in order to

reduce the MSE is not a feasible solution if non-response rate is supposed to be large enough.

The classical estimators such as T_{0HH}, T_{0R}^{*}, T_{0P}^{*}, discussed earlier in literature in presence of non-response are not helpful in the reduction of MSE to a desired level. In all these estimators, the only controlling factor for lowering down the MSE is L_i, if one desires so.

By utilizing FTE in order to propose separate- type estimators in the present work, we are able to control the precision of the estimator to a desired level only by making an appropriate choice of α.

Let the non-response rate and mean-square of the non-response group in the i^{th} stratum at a time be $W_{i2} = \frac{N_{i2}}{N_i}$ and S_{0i2}^2 respectively. Then, for a choice of $\alpha = \alpha_0$, the MSE of the estimator would be

$$M\left[T_{FS}(\alpha)/W_{i2}\right] = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)p_i^2\left[S_{0i}^2 + \phi(\alpha_0)^2 R_{01i}^2 S_{1i}^2 - 2\phi(\alpha_0)R_{01i}\rho_{01i}S_{0i}S_{1i}\right]$$

$$+ \sum_{i=1}^{k}\frac{L_i - 1}{n_i}W_{i2}p_i^2 S_{0i2}^2 \qquad (3.11)$$

Let us now suppose that the non-response rate increased over time and it is $W_{i2}^{'} = \frac{N_{i2}^{'}}{N_i}$ such that $N_{i2}^{'} > N_{i2}$. Obviously, with change in non-response rate, only the parameter S_{0i2}^2 will change. Let it becomes $S_{0i2}^{'2}$. Then we have

$$M\left[T_{FS}(\alpha)/W_{i2}^{'}\right] = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)p_i^2\left[S_{0i}^2 + \phi(\alpha_1)^2 R_{01i}^2 S_{1i}^2 - 2\phi(\alpha_1)R_{01i}\rho_{01i}S_{0i}S_{1i}\right]$$

$$+ \sum_{i=1}^{k}\frac{L_i - 1}{n_i}W_{i2}^{'}p_i^2 S_{0i2}^{'2} \qquad (3.12)$$

Clearly, if $\alpha_0 = \alpha_1$ and $S_{0i2}^{'2} > S_{0i2}^2$ then $M\left[T_{FS}(\alpha)\middle|W_{i2}^{'}\right] > M\left[T_{FS}(\alpha)\middle|W_{i2}\right]$. Therefore, we have to select a suitable value α_1, such that even if $W_{i2}^{'} > W_{i2}$ and $S_{0i2}^{'2} > S_{0i2}^2$, expression (3.12) becomes equal to equation (3.11) that is, the MSE of $T_{FS}(\alpha)$ is reduced to a desired level given by (3.11). Equating (3.11) to (3.12) and solving for $\phi(\alpha_1)$, we get

$$\phi(\alpha_1)^2 \sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2 - 2\phi(\alpha_1)\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}\rho_{01i}S_{0i}S_{1i}$$

$$-\left[\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2\left\{\phi(\alpha_0)^2 R_{01i}^2 S_{1i}^2 - 2\phi(\alpha_0)R_{01i}\rho_{01i}S_{0i}S_{1i}\right\}\right.$$

$$\left.+\sum_{i=1}^{k}\frac{L_i-1}{n_i}p_i^2\left(W_{i2}S_{0i2}^2 - W_{i2}^{'}S_{0i2}^{'2}\right)\right] = 0, \qquad (3.13)$$

which is quadratic equation in $\phi(\alpha_1)$. On solving the above equation, the roots are obtained as

$$\phi(\alpha_1) = \frac{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}\rho_{01i}S_{0i}S_{1i}}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2} \pm \left[\left\{\frac{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}\rho_{01i}S_{0i}S_{1i}}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2}\right\}^2 + \right.$$

$$\left.\frac{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2\left\{\phi(\alpha_0)^2 R_{01i}^2 S_{1i}^2 - 2\phi(\alpha_0)R_{01i}\rho_{01i}S_{0i}S_{1i}\right\} - \sum_{i=1}^{k}\frac{L_i-1}{n_i}p_i^2\left(W_{i2}^{'}S_{0i2}^{'2} - W_{i2}S_{0i2}^2\right)}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2}\right]^{\frac{1}{2}}$$

(3.14)

The above equation provides the value of α on which one can obtain the precision to a desired level. Sometimes the roots given by the above equation may be imaginary. So, in order that the roots are real, the conditions on the value of α_0 are given by

$$\phi(\alpha_0) > \frac{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}\rho_{01i}S_{0i}S_{1i}}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2} + \left[\frac{\sum_{i=1}^{k}\frac{L_i-1}{n_i}p_i^2\left(W_{i2}^{'}S_{0i2}^{'2}-W_{i2}S_{0i2}^2\right)}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2}\right]^{\frac{1}{2}} \qquad (3.15)$$

and $$\phi(\alpha_0) < \frac{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}\rho_{01i}S_{0i}S_{1i}}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2} - \left[\frac{\sum_{i=1}^{k}\frac{L_i-1}{n_i}p_i^2\left(W_{i2}^{'}S_{0i2}^{'2}-W_{i2}S_{0i2}^2\right)}{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2 R_{01i}^2 S_{1i}^2}\right]^{\frac{1}{2}} \qquad (3.16)$$

4. Empirical Study

In this section, therefore, we have illustrated the results, derived above, on the basis of some empirical data. For this purpose, a data set has been taken into consideration. Here the population is MU284 population available in Sarndal et. al. (1992, page 652, Appendix B). We have considered the population in the year 1985 as study variable and that in the year 1975 as auxiliary variable. There are 284 municipalities which have been divided randomly in to four strata having sizes 73, 70, 97 and 44.

Table 1 shows the values of the parameters of the population under consideration for the four strata which are needed in computational procedure.

Table 1: Parameters of the Population

Stratum (i)	**Stratum Size** (N_i)	**Mean** $(\overline{X}_{0i})$	**Mean** $(\overline{X}_{1i})$	(S_{0i}^2)	(S_{1i}^2)	S_{0i}	S_{1i}	ρ_{01i}	(S_{0i2}^2)
1	73	40.85	39.56	6369.0999	6624.4398	79.8066	81.3907	0.999	618.8844
2	70	27.83	27.57	1051.0725	1147.0111	32.4202	33.8676	0.998	240.9050
3	97	25.79	25.44	2014.9651	2205.4021	44.8884	46.9617	0.999	265.5220
4	44	20.64	20.36	538.4749	485.2655	23.2051	22.0287	0.997	83.6944

The value of $R_{01} = \overline{X}_0 / \overline{X}_1$ comes out to be 1.0192.

We fix the sample size to be 60. Then the allocation of samples to different strata under proportional and Neyman allocations are shown in the following table

Table 2: Allocation of Sample

Stratum (i)	**Size of Samples under**	
	Proportional Allocation	**Neyman Allocation**
1	15	26
2	15	10
3	21	19
4	9	5

On the basis of the equation (3.10), we obtained the optimum values of α:

Under Proportional Allocation

$\phi(\alpha)$= 0.9491, α_{opt} = (31.9975, 2.6128, 1.12) and

Under Neyman Allocation

$\phi(\alpha)$= 0.9527, α_{opt} = (34.1435, 2.6114, 1.1123).

The following table depicts the values of the MSE's of the estimators $T_{FS}(\alpha)$ for α_{opt}, $\alpha = 1$ and 4 under proportional and Neyman allocations. A comparison of MSE of $T_{FS}(\alpha)$ with α_{opt} and $\alpha = 1$ with that at $\alpha = 4$ reveals the fact that the utilization of auxiliary information at the estimation stage certainly improves the efficiency of the estimator as compared to the usual mean estimator T_{0st}^{*}.

Table 3: MSE Comparison ($L_i = 2$, $W_{i2} = 10\%$ for all i)

MSE	Allocation	
	Proportional	**Neyman**
$M[T_{FS}(\alpha)]_{\text{opt}}$	0.6264	0.6015
$M[T_{FS}(1)]$	0.7270	0.6705
$M[T_{FS}(4)]=V[T_{0st}^{*}]$	35.6069	28.6080

We shall now illustrate how by an appropriate choice of α, the MSE of the estimators $T_{FS}(\alpha)$ can be reduced to a desired level even if the non-response rate is increased.

Let us take $L_i = 2$, $W_{i2} = 0.1$, $W_{i2}^{'} = 0.3$ and $S_{0i2}^{'\ 2} = \frac{4}{3}\left(S_{0i2}^{2}\right)$ for all i

Under Proportional Allocation

From the condition (3.15) and (3.16), we have conditions for real roots of $\phi(\alpha_1)$ as

$\phi(\alpha_0) > 1.1527$ and $\phi(\alpha_0) < 0.7454$.

Therefore, if we take $\phi(\alpha_0) = 1.20$, then for this choice of $\phi(\alpha_0)$, we get

$$M[T_{FS}(\alpha)|W_{i2}] = 3.0712 \text{ and } M[T_{FS}(\alpha)|W_{i2}^{'}] = 4.6818.$$

Thus, there is about 52 percent increase in the MSE of the estimator if non-response rate is tripled. Now using (3.14), we get $\phi(\alpha_1)$=1.0957 and 0.8025. At this value of $\phi(\alpha_1)$, $M[T_{FS}(\alpha)]$ reduces to 3.0712 even if non-response rate is 30 percent. Thus a possible choice of α may be made in order to reduce the MSE to a desired level.

Under Neyman Allocation

Conditions for real roots of $\phi(\alpha_1)$

$\phi(\alpha_0) > 1.1746$ and $\phi(\alpha_0) < 0.7309$.

If $\phi(\alpha_0) = 1.20$ then we have

$M\left[T_{FS}(\alpha)|W_{i2}\right] = 2.4885$ and $M\left[T_{FS}(\alpha)|W'_{i2}\right] = 4.0072$.

Further, we get from (3.14), $\phi(\alpha_1)$=1.0620 and 0.8435, so that

$M\left[T_{FS}(\alpha)|W'_{i2}\right]$=2.4885 for $\phi(\alpha_1)$=1.0620.

5. Conclusion

We have suggested a general family of factor-type estimators for estimating the population mean in stratified random sampling under non-response using an auxiliary variable. The optimum property of the family has been discussed. It has also been discussed about the choice of appropriate estimator of the family in order to get a desired level of accuracy even if non-response is high. The Table 3 reveals that the optimum estimator of the suggested family has greater precision than separate ratio and sample mean estimators. Besides it, the reduction of MSE of the estimators $T_{FS}(\alpha)$ to a desired extent by an appropriate choice of the parameter α even if the non-response rate is high in the population, has also been illustrated.

References

Cochran, W. G. (1940): The estimation of the yields of cereal experiments by sampling for the ratio of grain in total produce. Journ. of The Agr. Sci., 30, 262-275.

Hansen, M. H. and Hurwitz, W. N. (1946): The problem of non-response in sample surveys. Journ. of The Amer. Statis. Assoc., 41, 517-529.

Hansen, M. H., Hurwitz, W. N. and Madow,, W. G. (1953): Sample Survey Methods and Theory, Volume Ii, John Wiley and Sons, Inc., New York.

Kadilar, C. and Cingi, H. (2003): Ratio estimators in stratified random sampling, Biom. Jour. 45 (2), 218–225.

Kadilar, C. and Cingi, H. (2005): A new ratio estimator in stratified sampling, Comm. in Stat.—Theor. and Meth., *34*, 597–602.

Khoshnevisan, M., Singh, R., Chauhan, P., Sawan, N., Smarandache, F. (2007): A general family of estimators for estimating population mean using known value of some population parameter(s), Far East Journ. of Theor. Statis., 22, 181–191.

Koyuncu, N. and Kadilar, C. (2008): Ratio and product estimators in stratified random sampling. Journ. of Statis. Plann. and Inf., 3820, 2-7.

Koyuncu, N. and Kadilar, C. (2009): Family of estimators of population mean using two auxiliary variables in stratified random sampling. Comm. in Stat.—Theor. and Meth., 38, 2398–2417.

Murthy, M. N. (1964): Product method of estimation. Sankhya, 26A, 69-74.

Sarndal, C. E., Swensson, B. and Wretman, J. (1992): Model Assisted Survey Sampling, Springer-Verlag, New York, Inc.

Searls, D. T. (1964): The utilization of a known coefficient of variation in the estimation procedure. Journ. of The Amer. Statis. Assoc., 59, 1225-1226.

Shabbir, J. and Gupta, S. (2005): Improved ratio estimators in stratified sampling. Amer. Journ. of Math. and Manag. Sci., 25 (3-4), 293-311.

Shabbir, J. and Gupta, S. (2006): A new estimator of population mean in stratified sampling. Comm. in Stat.—Theor. and Meth., 35, 1201–1209.

Singh, H. P., Tailor, R., Singh, S. and Kim, J. (2008): A modified estimator of population mean using power transformation. Stat. Paper., 49, 37-58.

Singh, H. P. and Vishwakarma, G. K. (2008): A Family of Estimators of Population Mean Using Auxiliary Information in Stratified Sampling**,** Comm. in Stat.—Theor. and Meth., 37, 1038–1050.

Singh, L. B. (1998): Some Classes of Estimators for Finite Population Mean in Presence of Non-response. Unpublished Ph. D. Thesis submitted to Banaras Hindu University, Varanasi, India.

Singh, V. K. and Shukla, D. (1987): One parameter family of factor-type ratio estimators. Metron, 45 (1-2), 273-283.

Srivenkataramana, T. (1980): A dual to ratio estimator in sample surveys, *Biometrika*, 67 (1), 199-204.

Upadhyaya, L. N. and Singh, H. P. (1999): Use of transformed auxiliary variable in estimating the finite population mean, Biom. Journ., 41 (5), 627-636.

On Some New Allocation Schemes in Stratified Random Sampling under Non-Response

Manoj Kr. Chaudhary, V. K. Singh and Rajesh Singh
Department of Statistics, Banaras Hindu University
Varanasi-221005, INDIA

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

Abstract

This chapter presents the detailed discussion on the effect of non-response on the estimator of population mean in a frequently used design, namely, stratified random sampling. In this chapter, our aim is to discuss the existing allocation schemes in presence of non-response and to suggest some new allocation schemes utilizing the knowledge of response and non-response rates of different strata. The effects of proposed schemes on the sampling variance of the estimator have been discussed and compared with the usual allocation schemes, namely, proportional allocation and Neyman allocation in presence of non-response. The empirical study has also been carried out in support of the results.

Keywords: Stratified random sampling, Allocation schemes, Non-response, Mean squares, Empirical Study.

1. Introduction

Sukhatme (1935) has shown that by effectively using the optimum allocation in stratified sampling, estimates of the strata variances obtained in a previous survey or in a specially planned pilot survey based even on samples of moderate sample size would be adequate for increasing the precision of the estimator. Evans (1951) has also considered the problem of allocation based on estimates of strata variances obtained in earlier survey. According to literature of sampling theory, various efforts have been made to reduce the error which arises because of taking a part of the population, *i.e.*, sampling

error. Besides the sampling error there are also several non-sampling errors which take place from time to time due to a number of factors such as faulty method of selection and estimation, incomplete coverage, difference in interviewers, lack of proper supervision, etc. Incompleteness or non-response in the form of absence, censoring or grouping is a troubling issue of many data sets.

In choosing the sample sizes from the different strata in stratified random sampling one can select it in such a way that it is either exclusively proportional to the strata sizes or proportional to strata sizes along with the variation in the strata under proportional allocation or Neyman allocation respectively. If non-response is inherent in the entire population and so are in all the strata, obviously it would be quite impossible to adopt Neyman allocation because then the knowledge of stratum variability will not be available, rather the knowledge of response rate of different strata might be easily available or might be easily estimated from the sample selected from each stratum. Thus, it is quite reasonable to utilize the response rate (or non-response rate) while allocating samples to stratum instead of Neyman allocation in presence of non-response error.

In the present chapter, we have proposed some new allocation schemes in selecting the samples from different strata based on response (non-response) rates of the strata in presence of non-response. We have compared them with Neyman and proportional allocations. The results have been shown with a numerical example.

2. Sampling Strategy and Estimation Procedure

In the study of non-response, according to one deterministic response model, it is generally assumed that the population is dichotomized in two strata; a response stratum considering of all units for which measurements would be obtained if the units happened to fall in the sample and a non-response stratum of units for which no measurement would be obtained. However, this division into two strata is, of course, an oversimplification of the problem. The theory involved in HH technique, is as given below:

Let us consider a sample of size n is drawn from a finite population of size N. Let n_1 units in the sample responded and n_2 units did not respond, so that $n_1 + n_2 = n$.

The n_1 units may be regarded as a sample from the response class and n_2 units as a sample from the non-response class belonging to the population. Let us assume that N_1 and N_2 be the number of units in the response stratum and non-response stratum respectively in the population. Obviously, N_1 and N_2 are not known but their unbiased estimates can be obtained from the sample as

$$\hat{N}_1 = n_1 N / n \,; \;\; \hat{N}_2 = n_2 N / n \,.$$

Let m be the size of the sub-sample from n_2 non-respondents to be interviewed. Hansen and Hurwitz (1946) proposed an estimator to estimate the population mean $\overline{X}_0$ of the study variable X_0 as

$$T_{0HH} = \frac{n_1 \bar{x}_{01} + n_2 \bar{x}_{0m}}{n}, \tag{2.1}$$

which is unbiased for $\overline{X}_0$, whereas $\bar{x}_{01}$ and $\bar{x}_{0m}$ are sample means based on samples of sizes n_1 and m respectively for the study variable X_0.

The variance of T_{0HH} is given by

$$V(T_{0HH}) = \left[\frac{1}{n} - \frac{1}{N}\right] S_0^2 + \frac{L-1}{n} W_2 S_{02}^2, \tag{2.2}$$

where $L = \frac{n_2}{m}$, $W_2 = \frac{N_2}{N}$, S_0^2 and S_{02}^2 are the mean squares of entire group and non-response group respectively in the population.

Let us consider a population consisting of N units divided into k strata. Let the size of i^{th} stratum is N_i, ($i = 1,2,...,k$) and we decide to select a sample of size n from the entire population in such a way that n_i units are selected from the i^{th} stratum. Thus, we have $\sum_{i=1}^{k} n_i = n$.

Let the non-response occurs in each stratum. Then using Hansen and Hurwitz procedure we select a sample of size m_i units out of n_{i2} non-respondent units in the i^{th} stratum with the help of simple random sampling without replacement (SRSWOR) such that $n_{i2} = L_i m_i$, $L_i \geq 1$ and the information are observed on all the m_i units by interview method.

The Hansen-Hurwitz estimator of population mean $\overline{X}_{0i}$ for the i^{th} stratum will be

$$T_{0i}^{*} = \frac{n_{i1}\bar{x}_{0i1} + n_{i2}\bar{x}_{0mi}}{n_i}, \qquad (i = 1,2,...,k) \tag{2.3}$$

where $\bar{x}_{0i1}$ and $\bar{x}_{0mi}$ are the sample means based on n_{i1} respondent units and m_i non-respondent units respectively in the i^{th} stratum.

Obviously T_{0i}^{*} is an unbiased estimator of $\overline{X}_{0i}$. Combining the estimators over all strata we get the estimator of population mean $\overline{X}_0$, given by

$$T_{0st}^{*} = \sum_{i=1}^{k} p_i T_{0i}^{*} \tag{2.4}$$

where $p_i = \frac{N_i}{N}$.

Obviously, we have

$$\mathrm{E}\left[T_{0st}^{*}\right] = \overline{X}_0. \tag{2.5}$$

The variance of T_{0st}^{*} is given by

$$V\left[T_{0st}^{*}\right] = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^{\,2} S_{0i}^{\,2} + \sum_{i=1}^{k} \frac{(L_i - 1)}{n_i} W_{i2}\, p_i^{\,2} S_{0i2}^{2} \tag{2.6}$$

where $W_{i2} = \frac{N_{i2}}{N_i}$, ${S_{0i}}^2$ and S_{0i2}^2 are the mean squares of entire group and non-response group respectively in the i^{th} stratum.

It is easy to see that under 'proportional allocation' (PA), that is, when $n_i = np_i$ for $i = 1,2,...,k$, $V\left[T_{0st}^*\right]$ is obtained as

$$V\left[T_{0st}^*\right]_{PA} = \sum_{i=1}^{k}\left(\frac{1}{n} - \frac{1}{N}\right)p_i {S_{0i}}^2 + \frac{1}{n}\sum_{i=1}^{k}(L_i - 1)W_{i2}p_i S_{0i2}^2, \tag{2.7}$$

whereas under the 'Neyman allocation' (NA), with $n_i = \frac{np_i S_{0i}}{\sum_{i=1}^{k} p_i S_{0i}}$ $\left(i = 1,2,...,k\right)$, it is equal to

$$V\left[T_{0st}^*\right]_{NA} = \frac{1}{n}\left(\sum_{i=1}^{k} p_i S_{0i}\right)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i {S_{0i}}^2 + \frac{1}{n}\left(\sum_{i=1}^{k}(L_i - 1)W_{i2}p_i \frac{S_{0i2}^2}{S_{0i}}\right)\left(\sum_{i=1}^{k} p_i S_{0i}\right). \tag{2.8}$$

It is important to mention here that the last terms in the expressions (2.7) and (2.8) arise due to non-response in the population. Further, in presence of non-response in the population, Neyman allocation may or may not be efficient than the proportional allocation, a situation which is quite contrary to the usual case when population is free from non-response. This can be understood from the following:

We have

$$V\left[T_{0st}^*\right]_{PA} - V\left[T_{0st}^*\right]_{NA} = \frac{1}{n}\sum_{i=1}^{k} p_i\left(S_{0i} - \overline{S}_w\right)^2 + \frac{1}{n}\sum_{i=1}^{k}(L_i - 1)p_i W_{i2} S_{0i2}^2\left(1 - \frac{\overline{S}_w}{S_{0i}}\right) \tag{2.9}$$

$$\overline{S}_w = \sum_{i=1}^{k} p_i S_{0i}\ .$$

Whole the first term in the above expression is necessarily positive, the second term may be negative and greater than the first term in magnitude depending upon the sign and magnitude of the term $\left(1-\frac{\overline{S}_w}{S_{0i}}\right)$ for all i. Thus, in presence of non-response in the stratified population, Neyman allocation does not always guarantee a better result as it is case when the population is free from non-response error.

3. Some New Allocation Schemes

It is a well known fact that in case the stratified population does not have non-response error and strata mean squares, $S_{0i}^2\ (i=1,2,...,k)$, are known, it is always advisable to prefer Neyman allocation scheme as compared to proportional allocation scheme in order to increase the precision of the estimator. But, if the population is affected by non-response, Neyman allocation is not always a better proposition. This has been highlighted under the section 2 above. Moreover, in case non-response is present in strata, knowledge on strata mean squares, S_{0i}^2, are impossible to collect, rather direct estimates of S_{0i1}^2 and S_{0i2}^2 may be had from the sample. Under these circumstances, it is, therefore, practically difficult to adopt Neyman allocation if non-response is inherent in the population. However, proportional allocation does not demand the knowledge of strata mean squares and rests only upon the strata sizes, hence it is well applicable even in the presence of non-response.

As discussed in the section 2, unbiased estimates of response and non-response rates in the population are readily available and hence it seems quite reasonable to think for developing allocation schemes which involve the knowledge of population response (non-response) rates in each stratum. If such allocation schemes yield précised estimates as compared to proportional allocation, these would be advisable to adopt instead of Neyman allocation due to the reasons mentioned above.

In this section, we have, therefore, proposed some new allocation schemes which utilize the knowledge of response (non-response) rates in subpopulations. While some of the proposed schemes do not utilize the knowledge of S_{0i}^2, some others are proposed

based on the knowledge of S_{0i}^2 just in order to make a comparison of them with Neyman allocation under the presence of non-response. In addition to the assumptions of proportional and Neyman allocations, we have further assume it logical to allocate larger sample from a stratum having larger number of respondents and vice-versa when proposing the new schemes of allocations.

Scheme-1[OA (1)]:

Let us assume that larger size sample is selected from a larger size stratum and with larger response rate, that is,

$$n_i \propto p_i W_{i1} \qquad \text{for} \qquad i = 1,2,...,k\,.$$

Then we have

$$n_i = K p_i W_{i1} \quad \text{where } K \text{ is a constant.}$$

The value of K will be

$$K = \frac{n}{\sum_{i=1}^{k} p_i W_{i1}}\,.$$

Thus we have

$$n_i = \frac{n p_i W_{i1}}{\sum_{i=1}^{k} p_i W_{i1}}\,. \tag{3.1}$$

Putting this value of n_i in expression (2.6), we get

$$V\left[T_{0st}^{*}\right]_1 = \frac{1}{n}\left[\sum_{i=1}^{k} p_i W_{i1}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i S_{0i}^2}{W_{i1}} + \frac{(L_i - 1)}{W_{i1}} W_{i2} p_i S_{0i2}^2\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 \tag{3.2}$$

Scheme-2[OA (2)]:

Let us assume that

$$n_i \propto p_i W_{i1} S_{0i} .$$

Then, we have

$$n_i = \frac{n p_i W_{i1} S_{0i}}{\sum_{i=1}^{k} p_i W_{i1} S_{0i}} \tag{3.3}$$

and hence the expression (2.6) becomes

$$V\left[T_{0st}^{*}\right]_2 = \frac{1}{n}\left[\sum_{i=1}^{k} p_i W_{i1} S_{0i}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i S_{0i}}{W_{i1}} + \frac{(L_i - 1) W_{i2} p_i}{W_{i1}} \frac{S_{0i2}^2}{S_{0i}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.4}$$

Scheme-3[OA (3)]:

Let us select larger size sample from a larger size stratum but smaller size sample if the non-response rate is high. That is,

$$n_i \propto \frac{p_i}{W_{i2}} .$$

Then

$$n_i = \frac{n p_i}{W_{i2} \sum_{i=1}^{k} \frac{p_i}{W_{i2}}} \tag{3.5}$$

and the expression of $V\left[T_{0st}^{*}\right]$ reduces to

$$V\left[T_{0st}^{*}\right]_3 = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{p_i}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{p_i W_{i2} S_{0i}^2 + (L_i - 1) W_{i2}^2 p_i S_{0i2}^2\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.6}$$

Scheme-4[OA (4)]:

Let

$$n_i \propto \frac{p_i S_{0i}}{W_{i2}}, \text{ then}$$

$$n_i = \frac{np_i S_{0i}}{W_{i2}\sum_{i=1}^{k}\frac{p_i S_{0i}}{W_{i2}}}. \tag{3.7}$$

The corresponding expression of $V\left[T_{0st}^{*}\right]$ is

$$V\left[T_{0st}^{*}\right]_4 = \frac{1}{n}\left[\sum_{i=1}^{k}\frac{p_i S_{0i}}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{p_i W_{i2} S_{0i} + (L_i - 1)W_{i2}^2 p_i \frac{S_{0i2}^2}{S_{0i}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.8}$$

Scheme-5[OA (5)]:

Let

$$n_i \propto \frac{p_i W_{i1}}{W_{i2}},$$

then

$$n_i = \frac{np_i W_{i1}}{W_{i2}\sum_{i=1}^{k}\frac{p_i W_{i1}}{W_{i2}}}. \tag{3.9}$$

The expression (2.6) gives

$$V\left[T_{0st}^{*}\right]_5 = \frac{1}{n}\left[\sum_{i=1}^{k}\frac{p_i W_{i1}}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i W_{i2} S_{0i}^2}{W_{i1}} + \frac{(L_i - 1)W_{i2}^2 p_i S_{0i2}^2}{W_{i1}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.10}$$

Scheme-6[OA (6)]:

If $n_i \propto \frac{p_i W_{i1} S_{0i}}{W_{i2}}$,

then, we have

$$n_i = \frac{n p_i W_{i1} S_{0i}}{W_{i2} \sum_{i=1}^{k} \frac{p_i W_{i1} S_{0i}}{W_{i2}}}. \tag{3.11}$$

In this case, $V\left[T_{0st}^{*}\right]$ becomes

$$V\left[T_{0st}^{*}\right]_6 = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{p_i W_{i1} S_{0i}}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i W_{i2} S_{0i}}{W_{i1}} + \frac{(L_i - 1) W_{i2}^2 p_i S_{0i2}^2}{W_{i1} S_{0i}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.12}$$

Remark 1: It is to be mentioned here that if response rate assumes same value in all the strata, that is $W_{i1} = W$ (say), then schemes 1, 3 and 5 reduces to 'proportional allocation', while the schemes 2, 4 and 6 reduces to 'Neyman allocation'. The corresponding expressions, $V\left[T_{0st}^{*}\right]_r$, $(r = 1,3,5)$ are then similar to $V\left[T_{0st}^{*}\right]_{PA}$ and $V\left[T_{0st}^{*}\right]_r$, $(r = 2,4,6)$ reduce to $V\left[T_{0st}^{*}\right]_{NA}$.

Remark 2: Although the theoretical comparison of expressions of $V\left[T_{0st}^{*}\right]_r$, $(r = 1,3,5)$ and $V\left[T_{0st}^{*}\right]_r$, $(r = 2,4,6)$ with $V\left[T_{0st}^{*}\right]_{PA}$ and $V\left[T_{0st}^{*}\right]_{NA}$ respectively is required in order to understand the suitability of the proposed schemes, but such comparisons do not yield explicit solutions in general. The suitability of a scheme does depend upon the parametric values of the population. We have, therefore, illustrated the results with the help of some empirical data.

4. Empirical Study

In order to investigate the efficiency of the estimator T_{0st}^{*} under proposed allocation schemes, based on response (non-response) rates, we have considered here an empirical data set.

We have taken the data available in Sarndal et. al. (1992) given in Appendix B. The data refer to 284 municipalities in Sweden, varying considerably in size and other characteristics. The population consisting of the 284 municipalities is referred to as the MU284 population.

For the purpose of illustration, we have randomly divided the 284 municipalities into four strata consisting of 73, 70, 97 and 44 municipalities. The 1985 population (in thousands) has been considered as the study variable, X_0.

On the basis of the data, the following values of parameters were obtained:

Table 1 : Particulars of the Data

($N = 284$)

Stratum (i)	**Size** (N_i)	**Stratum Mean** $(\overline{X}_{0i})$	**Stratum Mean Square** (S_{0i}^2)	**Mean Square of the Non-response Group** $(S_{0i2}^2) = \frac{4}{5} S_{0i}^2$
1	**73**	**40.85**	**6369.10**	**5095.28**
2	**70**	**27.83**	**1051.07**	**840.86**
3	**97**	**25.78**	**2014.97**	**1611.97**
4	**44**	**20.64**	**538.47**	**430.78**

We have taken sample size, $n = 60$.

Tables 2 depicts the values of sample sizes, n_i $(i=1,2,3,4)$ and values of $V\left[T_{0st}^{*}\right]$ under PA, NA and proposed schemes OA(1) to OA(6) for different selections of the values of L_i and W_{i2} $(i=1,2,3,4)$.

Table 2

Sample Sizes and Variance of T_{0st}^{*} under Different Allocation Schemes

(L_i =2.0, 2.5, 1.5, 3.5 for i = 1, 2, 3, 4 respectively)

Stratum	Non-response Rate (W_{i2}) (Percent)	Sample Size (n_i) and $V\left[T_{0st}^{*}\right]$ under															
		PA		NA		OA(1)		OA(2)		OA(3)		OA(4)		OA(5)		OA(6)	
		n_i	$V\left[T_{0st}^{*}\right]$	n_i	$V\left[T_{0st}^{*}\right]$	n_i	$V\left[T_{0st}^{*}\right]$	n_i	$V\left[T_{0st}^{*}\right]$	n_i	$V\left[T_{0st}^{*}\right]$	n_i	$V\left[T_{0st}^{*}\right]$	n_i	$V\left[T_{0st}^{*}\right]$	n_i	$V\left[T_{0st}^{*}\right]$
1	20	15	43.08	26	36.04	17	41.02	28	116.59	20	38.43	31	38.43	22	37.85	33	40.25
2	25	15		10		15		10		15		10		15		10	
3	30	21		19		20		18		18		16		17		14	
4	35	9		5		8		4		7		3		6		3	
1	35	15	45.97	26	37.27	14	49.17	24	117.37	12	55.41	21	39.07	10	60.76	19	40.72
2	30	15		10		14		10		13		10		13		10	
3	25	21		19		21		21		22		22		23		24	
4	20	9		5		10		5		13		7		14		7	
1	25	15	43.91	26	36.30	16	43.40	27	116.54	16	44.15	27	37.76	16	44.69	27	38.94
2	20	15		10		16		11		19		13		21		14	
3	30	21		19		20		18		18		17		17		16	
4	35	9		5		8		4		7		3		6		3	
1	20	15	43.17	26	35.99	17	41.32	28	115.40	20	39.45	32	38.82	22	39.73	34	41.30
2	25	15		10		15		10		16		10		16		10	
3	35	21		19		19		17		16		14		14		12	
4	30	9		5		9		5		8		4		8		4	

5. Concluding Remarks

In the present chapter, our aim was to accommodate the non-response error inherent in the stratified population during the estimation procedure and hence to suggest some new allocation schemes which utilize the knowledge of response (non-response) rates of strata. As discussed in different sub-sections, Neyman allocation may sometimes produce less précised estimates of population mean in comparison to proportional allocation if non-response is present in the population. Moreover, Neyman allocation is sometimes impractical in such situation, since then neither the knowledge of S_{0i} $(i=1,2,3,4)$, the mean squares of the strata, will be available, nor these could be estimated easily from the sample. In contrast to this, what might be easily known or could be estimated from the sample are response (non-response) rates of different strata. It was, therefore, thought to propose some new allocation schemes depending upon response (non-response) rates.

A look of Table 2 reveals that in most of the situations (under different combinations of W_{i2} and L_i), allocation schemes OA (1), OA (3) and OA (5), depending solely upon the knowledge of p_i and W_{i2} (or W_{i1}), produce more précised estimates as compared to PA. Further, as for as a comparative study of schemes OA (1), OA (3) and OA (5) is concerned, no doubt, all these schemes are more or less similar in terms of their efficiency. Thus, in addition to the knowledge of strata sizes, p_i, the knowledge of response (non-response) rates, W_{i1} (or W_{i2}), while allocating sample to different strata; certainly adds to the precision of the estimate.

It is also evident from the table that the additional information on the mean squares of strata certainly adds to the precision of the estimate, but this contribution is not very much significant in comparison to NA. Scheme OA (2) is throughout worse than any other scheme.

References

Evans, W. D. (1951) : On stratification and optimum allocation. Journ. of The Amer. Stat. Assoc., 46, 95-104.

Hansen, M. H. and Hurwitz, W. N. (1946) : The problem of non-response in sample Surveys. Journ. of The Amer. Stat. Assoc., 41, 517-529

Sarndal, C. E., Swensson, B. and Wretman, J. (1992) : Model Assisted Survey Sampling. Springer-Verlag, New York, Inc.

Sukhatme, P. V. (1935) : Contributions to the theory of the representative Method. Journ. of the Royal Stat. Soc., 2, 253-268.

A Family Of Estimators For Estimating The Population Mean In Stratified Sampling

Rajesh Singh, Mukesh Kumar, Sachin Malik and Manoj K. Chaudhary
Department of Statistics, B.H.U., Varanasi (U.P.)-India

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

Abstract

In this chapter, we have suggested an improved estimator for estimating the population mean in stratified sampling in presence of auxiliary information. The mean square error (MSE) of the proposed estimator have been derived under large sample approximation. Besides, considering the minimum case of the MSE equation, the efficient conditions between the proposed and existing estimators are obtained. These theoretical findings are supported by a numerical example.

Keywords : Auxiliary variable, mean square errors; exponential ratio type Estimates; stratified random sampling.

1. Introduction

In planning surveys, stratified sampling has often proved useful in improving the precision of other unstratified sampling strategies to estimate the finite population mean

$$\overline{Y} = \left(\sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}\right)/N$$

Consider a finite population of size N. Let y and x respectively, be the study and auxiliary variates on each unit U_j (j=1,2,3…N) of the population U. Let the population be divided in to L strata with the h^{th}stratum containing N_h units, h=1,2,3…,L so that

$\sum_{h=1}^{L} N_h = N$. Suppose that a simple random sample of size n_h is drawn without replacement (SRSWOR) from the h^{th} stratum such that $\sum_{h=1}^{L} n_h = n$.

When the population mean $\bar{X}$ of the auxiliary variable x is known, Hansen et. al. (1946) suggested a "combined ratio estimator "

$$\bar{y}_{CR} = \bar{y}_{st}\left(\frac{\bar{X}}{\bar{x}_{st}}\right) \tag{1.1}$$

where, $\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h$, $\bar{x}_{st} = \sum_{h=1}^{L} w_h \bar{x}_h$

$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ and $\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$

$w_h = \frac{N_h}{N}$ and $\bar{X} = \sum_{h=1}^{L} w_h \bar{X}_h$.

The "combined product estimator " for $\bar{Y}$ is defined by

$$\bar{y}_{CP} = \bar{y}_{st}\left(\frac{\bar{x}_{st}}{\bar{X}}\right) \tag{1.2}$$

To the first degree of approximation, the mean square error (MSE) of $\bar{y}_{CR}$ and $\bar{y}_{CP}$ are respectively given by –

$$MSE(\bar{y}_{CR}) \cong \sum_{i=1}^{L} w_h^2 \theta_h \left[S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{yxh}\right] \tag{1.3}$$

$$MSE(\bar{y}_{CP}) \cong \sum_{i=1}^{L} w_h^2 \theta_h \left[S_{yh}^2 + R^2 S_{xh}^2 + 2RS_{yxh}\right] \tag{1.4}$$

where $\theta_h = \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$, $R = \frac{\bar{Y}}{\bar{X}}$ is the population ratio, S_{yh}^2 is the population variance of variate of interest in stratum h, S_{xh}^2 is the population variance of auxiliary variate in stratum h and S_{yxh} is the population covariance between auxiliary variate and variate of interest in stratum h.

Following Bahl and Tuteja (1991), Singh et. al. (2009) proposed following estimator in stratified random sampling -

$$\bar{y}_{er} = \bar{y}_{st} \exp\left[\frac{\bar{X} - \bar{x}_{st}}{\bar{X} + \bar{x}_{st}}\right] \tag{1.5}$$

The MSE of $\bar{y}_{er}$, to the first degree of approximation is given by

$$MSE(\bar{y}_{er}) \cong \sum_{i=1}^{L} w_h^2 \theta_h \left[S_{yh}^2 + \frac{R^2}{4} S_{xh}^2 - RS_{yxh} \right] \quad (1.6)$$

Using the estimator $\bar{y}_{CR}$ and $\bar{y}_{CP}$, Singh and Vishwakarma (2005) suggested the combined ratio-product estimator for estimating $\bar{Y}$ as

$$\bar{y}_{RPC} = \bar{y}_{st} \left[\alpha \frac{\bar{X}}{\bar{x}_{st}} + (1 - \alpha) \frac{\bar{x}_{st}}{\bar{X}} \right] \quad (1.7)$$

For minimum value of $\alpha = \frac{1}{2}(1 + C^*) = \alpha_0$ (say), the minimum MSE of the estimator $\bar{y}_{RPC}$ is given by

$$MSE\ (\bar{y}_{RPC}) = \sum_{i=1}^{L} w_h^2 \theta_h (1 - \rho^{*2})\ S_{yh}^2 \quad (1.8)$$

where $C^* = \frac{cov(\bar{y}_{st}, \bar{x}_{st})}{RV(\bar{x}_{st})}$, $\rho^* = \frac{cov(\bar{y}_{st}, \bar{x}_{st})}{R\sqrt{V(\bar{y}_{st})V(\bar{x}_{st})}}$, $R = \frac{\bar{Y}}{\bar{X}}$.

2. Proposed estimator

Following Singh and Vishwakarma (2005), we propose a new family of estimators -

$$t = \lambda \left[\bar{y}_{st} \exp \left[\frac{\bar{X} - \bar{x}_{st}}{\bar{X} + \bar{x}_{st}} \right]^{\alpha} \left(\frac{\bar{X}}{\bar{x}_{st}} \right)^{\beta} \right] + (1 - \lambda) \left[\bar{y}_{st} \exp \left[\frac{\bar{x}_{st} - \bar{X}}{\bar{x}_{st} + \bar{X}} \right]^{\alpha} \left(\frac{\bar{x}_{st}}{\bar{X}} \right)^{\beta} \right] \quad (2.1)$$

where λ is real constant to be determined such that the MSE of t is a minimum and α, β are real constants such that β =1- α.

Remark 2.1: For $\lambda = 1$ and $\alpha = 1$ the estimator t tends to Singh et. al. (2009) estimator. For $\lambda = 1$ and $\alpha = 0$ the estimator t takes the form of Hansen et. al. (1946) estimator $\bar{y}_{CR}$. For $\lambda = 0$ and $\alpha = 1$ the estimator t tends to Singh et. al. (2009) estimator. For $\lambda = 1$ and $\alpha = 0$ the estimator t takes the form of the estimator $\bar{y}_{CP}$.

To obtain the MSE of t to the first degree of approximation, we write

$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h = \bar{Y}(1 + e_0)$ and

$\bar{x}_{st} = \sum_{h=1}^{L} w_h \bar{x}_h = \bar{X}(1 + e_1)$

Such that,

$E(e_0) = E(e_0) = 0.$

Under SRSWOR, we have

$E(e_0^2) = \frac{1}{\bar{Y}^2} \sum_{i=1}^{L} w_h^2 \theta_h S_{yh}^2$

$E(e_1^2) = \frac{1}{\bar{X}^2} \sum_{i=1}^{L} w_h^2 \theta_h S_{xh}^2$

$E(e_0 e_1) = \frac{1}{\bar{Y}\bar{X}} \sum_{i=1}^{L} w_h^2 \theta_h S_{yxh}$

Expressing equation (2.1) in terms of e's we have

$$t = \bar{Y}(1 + e_0)\left[\lambda\left\{\exp\left(\frac{-e_1}{2}\left(1 + \frac{e_1}{2}\right)^{-1}\right)\right\}^{\alpha}\{(1 + e_1)^{-1}\}^{1-\alpha} + (1 - \lambda)\left\{\exp\left(\frac{e_1}{2}\left(1 + \frac{e_1}{2}\right)^{-1}\right)\right\}^{1-\alpha}(1 + e_1)^{(1-\alpha)}\right] \quad (2.2)$$

We now assume that $|e_1| < 1$ so that we may expand $(1 + e_1)^{-1}$ as a series in powers of e_1. Expanding the right hand side of (2.2) to the first order of approximation, we obtain

$$(t - \bar{Y}) \cong \bar{Y}\left[e_0 + e_1\left(1 + \alpha\lambda - \frac{\alpha}{2} - 2\lambda\right)\right] \quad (2.3)$$

Squaring both sides of (2.3) and then taking expectations, we get the MSE of the estimator t, to the first order of approximation, as

$$MSE(t) = V(\bar{y}_{st}) + R^2(1 - 2\lambda)S_{xh}^2\{(1 - 2\lambda)A^2 + 2C^*A\} \quad (2.4)$$

where $A = \left(1 - \frac{\alpha}{2}\right)$.

Minimisation of (2.4) with respect to λ yields its optimum values as

$$\lambda_{opt} = \frac{1}{2}\left(1 + \frac{C^*}{A}\right) = \lambda_0(\text{say}) \quad (2.5)$$

Putting $\lambda = \lambda_0$ in (2.4) we get the minimum MSE of the estimator t as –

$$\min \mathrm{MSE}\,(t) = V(\bar{y}_{st})\,(1-\rho^{*2})$$

$$= \sum_{i=1}^{L} w_h^2 \theta_h (1-\rho^{*2}) S_{yh}^2. \qquad (2.6)$$

3. Efficiency comparisons

In this section we have compared proposed estimator with different already proposed estimators, obtained the conditions under which our proposed estimator performs better than other estimators.

First we have compared proposed estimator with simple mean in stratified random sampling.

$\mathrm{MSE}(t) \le \mathrm{MSE}(\bar{y}_{st})$, if

$V(\bar{y}_{st}) + R^2(1-2\lambda)S_{xh}^2\{(1-2\lambda)A^2 + 2C^*A\} \le V(\bar{y}_{st})$

$\min\left(\frac{1}{2} - \frac{1}{2} + \frac{C^*}{4}\right) \le \lambda \le \max\left(\frac{1}{2} - \frac{1}{2} + \frac{C^*}{4}\right)$

Next we compare proposed estimator with combined ratio estimator –

$\mathrm{MSE}(t) \le \mathrm{MSE}(\bar{y}_{CR})$, if

$V(\bar{y}_{st}) + \sum_{i=1}^{L} w_h^2 \theta_h\, R^2(1-2\lambda)S_{xh}^2\{(1-2\lambda)A^2 + 2C^*A\} \le$

$$\sum_{i=1}^{L} w_h^2 \theta_h \left[S_{yh}^2 + R^2 S_{xh}^2 - 2RS_{yxh}\right]$$

or, if $(1-2C^*) - (1-2\lambda)\left((1-2\lambda)A^2 + 2C^*A\right) \ge 0$

or, if $\frac{1}{2}\left\{\frac{A+1}{A}\right\} \le \lambda \le \frac{1}{2}\left\{\frac{2C^*+A-1}{A}\right\}$.

Next we compare efficiency of proposed estimator with product estimator

$\mathrm{MSE}(t) \le \mathrm{MSE}(\bar{y}_{PR})$, if

$V(\bar{y}_{st}) + \sum_{i=1}^{L} w_h^2 \theta_h\, R^2(1-2\lambda)S_{xh}^2\{(1-2\lambda)A^2 + 2C^*A\} \le$

$$\sum_{i=1}^{L} w_h^2 \theta_h \left[S_{yh}^2 + R^2 S_{xh}^2 + 2RS_{yxh}\right]$$

or, if $(1+2C^*)-(1-2\lambda)((1-2\lambda)A^2+2C^*A)\geq 0$

or, if $\frac{1}{2}\left\{\frac{A-1}{A}\right\}\leq \lambda \leq \frac{1}{2}\left\{\frac{2C^*+A+1}{A}\right\}$.

Next we compare efficiency of proposed estimator and exponential ratio estimator in stratified sampling

$MSE(t)\leq MSE(\bar{y}_{ER})$, if

$V(\bar{y}_{st})+\sum_{i=1}^{L} w_h^2\theta_h R^2(1-2\lambda)S_{xh}^2\{(1-2\lambda)A^2+2C^*A\}\leq$

$$\sum_{i=1}^{L} w_h^2\theta_h\left[S_{yh}^2+\frac{R^2}{4}S_{xh}^2-RS_{yxh}\right]$$

or, if $(1-4C^*)-4(1-2\lambda)((1-2\lambda)A^2+2C^*A)\geq 0$

or, if $\frac{1}{2}\left\{\frac{1-2A}{2A}\right\}\leq \lambda \leq \frac{1}{2}\left\{\frac{4C^*+2A-1}{2A}\right\}$

Finally we compare efficiency of proposed estimator with exponential product estimator in stratified random sampling

$MSE(t)\leq MSE(\bar{y}_{EP})$, if

or, if $V(\bar{y}_{st})+\sum_{i=1}^{L} w_h^2\theta_h R^2(1-2\lambda)S_{xh}^2\{(1-2\lambda)A^2+2C^*A\}\leq$

$$\sum_{i=1}^{L} w_h^2\theta_h\left[S_{yh}^2+\frac{R^2}{4}S_{xh}^2+RS_{yxh}\right]$$

or, if $(1+4C^*)-4(1-2\lambda)((1-2\lambda)A^2+2C^*A)\geq 0$

or, if $\frac{1}{2}\left\{\frac{-1-2A}{2A}\right\}\leq \lambda \leq \frac{1}{2}\left\{\frac{4C^*+2A+1}{2A}\right\}$

Whenever above conditions are satisfied the proposed estimator performs better than other mentioned estimators.

4. Numerical illustration

All the theoretical results are supported by using the data given in Singh and Vishwakarma (2005).

Data statistics:

Stratum	$\mathbf{w_h}$	$\boldsymbol{\theta_h}$	$\mathbf{S_{xh}^2}$	$\mathbf{S_{yh}^2}$	$\mathbf{S_{yxh}}$
2	0.5227	0.12454	132.66	259113.71	5709.16
3	0.2428	0.08902	38.44	65885.60	1404.71

R=49.03 and $\lambda_{opt} = 0.9422(\alpha = 0)$ and $1.384525\ (\alpha = 1)$

Using the above data percentage relative efficiencies of different estimators $\bar{y}_{CR}$, $\bar{y}_{CP}$, $\bar{y}_{ER}$, $\bar{y}_{EP}$ and proposed estimator t w.r.t $\bar{y}_{st}$ have been calculated.

Table 4.1: PRE of different estimators of $\bar{Y}$

Estimator	$\bar{y}_{st}$	$\bar{y}_{CR}$	$\bar{y}_{CP}$	$\bar{y}_{ER}$	$\bar{y}_{EP}$	$\bar{y}_{HPS(opt)}$	$\bar{y}_{PRP(opt)}$
PRE	100	1148.256	23.326	405.222	42.612	1403.317	1403.317

We have also shown the range of λ for which proposed estimator performs better than $\bar{y}_{st}$.

Table 4.2: Range of λ for which proposed estimator performs better than $\bar{y}_{st}$

Value of constant α	Form of proposed estimator	Range of λ
$\alpha = 0$	$\bar{y}_{HPS}$	(0.5,1.3)
$\alpha = 1$	$\bar{y}_{CER}$	(0.5,2.2)

5. Conclusion

From the theoretical discussion and empirical study we conclude that the proposed estimator under optimum conditions performs better than other estimators considered in the article. The relative efficiency of various estimators are listed in Table

4.1 and the range of λ for which proposed estimator performs better than $\bar{y}_{st}$ is written in Table 4.2.

References

Bahl. S. and Tuteja, R.K. (1991): Ratio and Product Type Exponential Estimator. Infrm. and Optim. Sci., XIII, 159-163.

Hansen, M.H. and Hurwitz, W.N. (1946): The problem of non-response in sample surveys. J. Am. Stat. Assoc. 41:517–529.

Singh, H. P. and Vishwakarma, G. K. (2005): Combined Ratio-Product Estimator of Finite Population Mean in Stratified Sampling. *Metodologia de Encuestas* 8: 35-44.

Singh, R., Kumar, M., Chaudhary, M.K., Kadilar, C. (2009) : Improved Exponential Estimator in Stratified Random Sampling. Pak. J. Stat. Oper. Res. 5(2), pp 67-82.

A Family Of Estimators Of Population Variance Using Information On Auxiliary Attribute

Rajesh Singh and Mukesh Kumar
Department of Statistics, B.H.U., Varanasi (U.P.)-India

Ashish K. Singh
College of Management Studies,
Raj Kumar Goel Institute of Technology

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

Abstract

This chapter proposes some estimators for the population variance of the variable under study, which make use of information regarding the population proportion possessing certain attribute. Under simple random sampling without replacement (SRSWOR) scheme, the mean squared error (MSE) up to the first order of approximation is derived. The results have been illustrated numerically by taking some empirical population considered in the literature.

Keywords: Auxiliary attribute, exponential ratio-type estimates, simple random sampling, mean square error, efficiency.

1. Introduction

It is well known that the auxiliary information in the theory of sampling is used to increase the efficiency of estimator of population parameters. Out of many ratio, regression and product methods of estimation are good examples in this context. There exist situations when information is available in the form of attribute which is highly correlated with y. Taking into consideration the point biserial correlation coefficient between auxiliary attribute and study variable, several authors including Naik and

Gupta (1996), Jhajj et. al. (2006), Shabbir and Gupta (2007), Singh et. al. (2007, 2008) and Abd-Elfattah et. al. (2010) defined ratio estimators of population mean when the prior information of population proportion of units, possessing the same attribute is available.

In many situations, the problem of estimating the population variance σ^2 of study variable y assumes importance. When the prior information on parameters of auxiliary variable(s) is available, Das and Tripathi (1978), Isaki (1983), Prasad and Singh (1990), Kadilar and Cingi (2006, 2007) and Singh et. al. (2007) have suggested various estimators of S_y^2.

In this chapter we have proposed family of estimators for the population variance S_y^2 when one of the variables is in the form of attribute. For main results we confine ourselves to sampling scheme SRSWOR ignoring the finite population correction.

2. The proposed estimators and their properties

Following Isaki (1983), we propose a ratio estimator

$$t_1 = s_y^2 \frac{S_\phi^2}{s_\phi^2} \tag{2.1}$$

Next we propose regression estimator for the population variance

$$t_2 = s_y^2 + b\left(S_\phi^2 - s_\phi^2\right) \tag{2.2}$$

And following Singh et. al. (2009), we propose another estimator

$$t_3 = s_y^2 \exp\left[\frac{S_\phi^2 - s_\phi^2}{S_\phi^2 - s_\phi^2}\right] \tag{2.3}$$

where s_y^2 and s_ϕ^2 are unbiased estimator of population variances S_y^2 and S_ϕ^2 respectively and b is a constant, which makes the MSE of the estimator minimum.

To obtain the bias and MSE, we write-

$$s_y^2 = S_y^2(1+e_0), \qquad s_\phi^2 = S_\phi^2(1+e_1)$$

Such that $E(e_0) = E(e_1) = 0$

and $E(e_0^2) = \frac{(\delta_{40}-1)}{n}, \qquad E(e_1^2) = \frac{(\delta_{04}-1)}{n}, \qquad E(e_0 e_1) = \frac{(\delta_{22}-1)}{n},$

where $\delta_{pq} = \frac{\mu_{pq}}{(\mu_{20}^{p/2}\mu_{02}^{q/2})}, \quad \mu_{pq} = \frac{\sum_{i=1}^{N}(y_i - \overline{Y})^p(\phi_i - P)^q}{(N-1)}$.

$\beta_{2(y)} = \frac{\mu_{40}}{\mu_{02}^2} = \delta_{40}$ and $\beta_{2(\phi)} = \frac{\mu_{04}}{\mu_{02}^2} = \delta_{04}$

Let $\beta_{2(y)}^* = \beta_{2(y)} - 1, \beta_{2(\phi)}^* = \beta_{2(x)} - 1,$ and $\delta_{pq}^* = \delta_{pq} - 1$

P is the proportions of units in the population.

Now the estimator t_1 defined in (2.1) can be written as

$$(t_1 - S_y^2) = S_y^2(e_0 - e_1 + e_1^2 - e_0 e_1) \tag{2.4}$$

Similarly, the estimator t_2 can be written as

$$(t_2 - S_y^2) = S_y^2 e_0 - bS_\phi^2 e_1 \tag{2.5}$$

And the estimator t_3 can be written as

$$(t_3 - S_y^2) = S_y^2\left(e_0 - \frac{e_1}{2} - \frac{e_0 e_1}{2} + \frac{3e_1^2}{8}\right) \tag{2.6}$$

The MSE of t_1, t_3 and variance of t_2 are given, respectively, as

$$MSE(t_{p1}) = \frac{S_y^4}{n}\left[\beta^*_{2(y)} + \beta^*_{2(\phi)} - 2\delta^*_{22}\right] \tag{2.7}$$

$$MSE(t_{p3}) = \frac{S_y^4}{n}\left[\beta^*_{2(y)} + \frac{\beta^*_{2(\phi)}}{4} - \delta^*_{22}\right] \tag{2.8}$$

The variance of t_{p2} is given as

$$V(t_2) = \frac{1}{n}\left[S_y^4(\lambda_{40}-1) + b^2 S_\phi^2(\lambda_{04}-1) - 2bS_y^2 S_x^2(\lambda_{22}-1)\right] \tag{2.9}$$

On differentiating (2.9) with respect to b and equating to zero we obtain

$$b = \frac{S_y^2(\delta_{22}-1)}{S_x^2(\delta_{04}-1)} \tag{2.10}$$

Substituting the optimum value of b in (2.9), we get the minimum variance of the estimator t_2, as

$$\min.V(t_2) = \frac{S_y^4}{n}\beta^*_{2(y)}\left[1 - \frac{\delta^{*2}_{22}}{\beta^*_{2(y)}\beta^*_{2(\phi)}}\right] = Var(\bar{S}^2)\left(1 - \rho^2_{(S_y^2, S_\phi^2)}\right) \tag{2.11}$$

3. Adapted estimator

We adapt the Shabbir and Gupta (2007) and Grover (2010) estimator, to the case when one of the variables is in the form of attribute and propose the estimator t_4

$$t_4 = \left[k_1 s_y^2 + k_2\left(S_\phi^2 - s_\phi^2\right)\right]\exp\left(\frac{S_\phi^2 - s_\phi^2}{S_\phi^2 + s_\phi^2}\right) \tag{3.1}$$

where k_1 and k_2 are suitably chosen constants.

Expressing equation (3.1) in terms of e's and retaining only terms up to second degree of e's, we have:

$$t_4 = \left[k_1 S_y^2(1+e_0) - k_2 s_\phi^2 e_1\right]\left[1 - \frac{e_1}{2} + \frac{3}{8}e_1^2\right] \tag{3.2}$$

Up to first order of approximation, the mean square error of t_4 is

$$\begin{aligned} MSE(t_4) &= E\left(t_4 - S_y^2\right)^2 \\ &= S_y^4\left[(k_1-1)^2 + \lambda k_1^2\left(\beta_2^*(y)+\beta_2^*(\phi)-2\delta_{22}\right)+\lambda k_1\left(\delta_{22}^* - \frac{3}{4}\beta_2^*(\phi)\right)\right. \\ &\left. + S_\phi^4 k_2^2\lambda\beta_2^*(\phi) + 2\lambda S_y^2 S_x^2\left[k_1 k_2\left(\beta_2^*(x)-\delta_{22}^*\right)-\frac{k_2}{2}\beta_2^*(x)\right]\right] \end{aligned} \tag{3.3}$$

where, $\lambda = \frac{1}{n}$

On partially differentiating (3.3) with respect to $k_i (i = 1,2)$, we get optimum values of k_1 and k_2, respectively as

$$k_1^* = \frac{\beta_2^*(\phi)\left(2-\frac{\lambda}{4}\beta_2^*(\phi)\right)}{2\left(\beta_2^*(\phi)(\lambda A+1)-\lambda B^2\right)} \tag{3.4}$$

and

$$k_2^* = \frac{S_y^2\left[\beta_2^*(\phi)(\lambda A+1)-\lambda B^2 - B\left(2-\frac{\lambda}{4}\beta_2^*(\phi)\right)\right]}{2S_x^2\left(\beta_2^x(\phi)(\lambda A+1)-\lambda B^2\right)} \tag{3.5}$$

where,

$A = \beta_2^*(y)+\beta_2^*(\phi)-2\delta_{22}^*$ and $B = \beta_2^*(\phi)-\delta_{22}^*$.

On substituting these optimum values of k_1 and k_2 in (3.3), we get the minimum value of MSE of t_4 as

$$MSE(t_4) = \frac{MSE(t_2)}{1+\frac{MSE(t_2)}{S_y^4}} - \frac{\lambda\beta_2^*(x)\left(MSE(t_2)+\frac{\lambda S_y^4\beta_2^*(\phi)}{16}\right)}{4\left(1+\frac{MSE(t_2)}{S_y^4}\right)} \tag{3.6}$$

4. Efficiency Comparison

First we have compared the efficiency of proposed estimator under optimum condition with the usual estimator as -

$$V\left(\hat{S}_y^2\right) - MSE\left(\hat{S}_p^2\right)_{opt} = \frac{\lambda S_y^4 \delta_{22}^{*2}}{\beta_{2(x)}^*} - \frac{MSE(t_2)}{1 + \frac{MSE(t_2)}{S_y^4}} + \frac{\lambda \beta_2^*(x)\left(MSE(t_2) + \frac{\lambda S_y^4 \beta_2^*(\phi)}{16}\right)}{4\left(1 + \frac{MSE(t_2)}{S_y^4}\right)} \geq 0 \text{ always.} \quad (4.1)$$

Next we have compared the efficiency of proposed estimator under optimum condition with the ratio estimator as -

From (2.1) and (3.6) we have

$$MSE(t_2) - MSE\left(\hat{S}_p^2\right)_{opt} = \lambda S_y^4 \left[\sqrt{\beta_{2(x)}} - \frac{\delta_{22}^*}{\sqrt{\beta_{2(x)}^*}}\right]^2 - \frac{MSE(t_2)}{1 + \frac{MSE(t_2)}{S_y^4}} + \frac{\lambda \beta_2^*(x)\left(MSE(t_2) + \frac{\lambda S_y^4 \beta_2^*(\phi)}{16}\right)}{4\left(1 + \frac{MSE(t_2)}{S_y^4}\right)} \geq 0 \text{ always.} \quad (4.2)$$

Next we have compared the efficiency of proposed estimator under optimum condition with the exponential ratio estimator as -

From (2.3) and (3.6) we have

$$MSE(t_3) - MSE\left(\hat{S}_p^2\right)_{opt} = \lambda S_y^4 \left[\sqrt{\beta_{2(x)}} - \frac{\delta_{22}^*}{2\sqrt{\beta_{2(x)}^*}} \right]^2 - \frac{MSE(t_2)}{1 + \frac{MSE(t_2)}{S_y^4}}$$

$$+ \frac{\lambda \beta_2^*(x)\left(MSE(t_2) + \frac{\lambda S_y^4 \beta_2^*(\phi)}{16} \right)}{4\left(1 + \frac{MSE(t_2)}{S_y^4} \right)} \geq 0 \text{ always.} \qquad (4.3)$$

Finally we have compared the efficiency of proposed estimator under optimum condition with the Regression estimator as -

$$MSE(t_2) - MSE(t_4) = \frac{MSE(t_2)}{1 + \frac{MSE(t_2)}{S_y^4}} - \frac{\lambda \beta_2^*(x)\left(MSE(t_2) + \frac{\lambda S_y^4 \beta_2^*(\phi)}{16} \right)}{4\left(1 + \frac{MSE(t_2)}{S_y^4} \right)} > 0 \text{ always.} \qquad (4.4)$$

5. Empirical study

We have used the data given in Sukhatme and Sukhatme (1970), p. 256.
Where, Y=Number of villages in the circle, and
ϕ Represent a circle consisting more than five villages.

n	N	S_y^2	S_p^2	λ_{40}	λ_{04}	λ_{22}
23	89	4.074	0.110	3.811	6.162	3.996

The following table shows PRE of different estimator's w. r. t. to usual estimator.

Table 1: PRE of different estimators

Estimators	t_0	t_1	t_2	t_3	t_4
PRE	100	141.898	262.187	254.274	**296.016**

Conclusion

Superiority of the proposed estimator is established theoretically by the universally true conditions derived in Sections 4. Results in Table 1 confirms this superiority numerically using the previously used data set.

References

Abd-Elfattah, A.M. El-Sherpieny, E.A. Mohamed, S.M. Abdou, O. F. (2010): Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute. Appl. Mathe. and Compt.

Das, A. K., Tripathi, T. P. (1978). Use of auxiliary information in estimating the finite population variance. Sankhya 40:139–148.

Grover, L.K. (2010): A Correction Note on Improvement in Variance Estimation Using Auxiliary Information. Communications in Statistics—Theory and Methods, 39: 753–764, 2010

Kadilar, C., Cingi, H. (2006). Improvement in variance estimation using auxiliary information. Hacettepe J. Math. Statist. 35(1):111–115.

Kadilar, C., Cingi, H. (2007). Improvement in variance estimation in simple random sampling. Commun. Statist. Theor. Meth. 36:2075–2081.

Isaki, C. T. (1983). Variance estimation using auxiliary information, Jour. of Amer. Statist.Asso.78, 117–123, 1983.

Jhajj, H.S., Sharma, M.K. and Grover, L.K. (2006). A family of estimators of Population mean using information on auxiliary attribute. Pak. J. Statist., 22(1),43-50.

Naik, V.D., Gupta, P.C. (1996): A note on estimation of mean with known population of an auxiliary character, Journal of Ind. Soci. Agri. Statist. 48(2) 151–158.

Prasad, B., Singh, H. P. (1990). Some improved ratio-type estimators of finite population variance in sample surveys. Commun. Statist. Theor. Meth. 19:1127–1139

Singh, R. Chauhan, P. Sawan, N. Smarandache, F. (2007): A general family of estimators for estimating population variance using known value of some population parameter(s). Renaissance High Press.

Singh, R. Chauhan, P. Sawan, N. Smarandache, F. (2008): Ratio estimators in simple random sampling using information on auxiliary attribute. Pak. J. Stat. Oper. Res. 4(1) 47–53

Shabbir, J., Gupta, S. (2007): On estimating the finite population mean with known population proportion of an auxiliary variable. Pak. Jour. of Statist.
23 (1) 1–9.

Shabbir, J., Gupta, S. (2007). On improvement in variance estimation using auxiliary information. Commun. Statist. Theor. Meth. 36(12):2177–2185.

	Observed Data	*Trend values*	*Single exponential smoothing (alpha=0.1)*	*Double exponential smoothing (alpha=0.9 and gamma=0.1)*
1995	*5.12*	*5.3372*		*5.12*
1996	*5.75*	*5.3036*	*5.12*	*5.707*
1997	*5.26*	*5.27*	*5.183*	*5.32857*
1998	*5.72*	*5.2364*	*5.1907*	*5.698556*
1999	*4.64*	*5.2028*	*5.24363*	*4.765484*
2000	*5.14*	*5.1692*	*5.183267*	*5.110884*
2003	*4.23*	*5.1356*	*5.17894*	*4.329044*
2004	*6.026*	*5.102*	*5.084046*	*5.858346*
2005	*4.46*	*5.0684*	*5.178242*	*4.616965*
2006	*5.52*	*5.0348*	*5.106417*	*5.4327*
Total	*51.866*	*51.86*	*46.46824*	*51.96755*

This book has been designed for students and researchers who are working in the field of time series analysis and estimation in finite population. There are papers by Rajesh Singh, Florentin Smarandache, Shweta Maurya, Ashish K. Singh, Manoj Kr. Chaudhary, V. K. Singh, Mukesh Kumar and Sachin Malik.

First chapter deals with the problem of time series analysis and the rest of four chapters deal with the problems of estimation in finite population.

ISBN 9781599731599